Textbook English: A Corpus-Based Analysis of the Language of EFL Textbooks used in Secondary Schools in France, Germany and Spain

Elen Le Foll

UNIVERSITÄT OSNABRÜCK

Dissertation

zur Erlangung des Doktorgrades (Dr. phil.)
des Fachbereichs Sprach- und Literaturwissenschaft
der Universität Osnabrück

zum Thema

Textbook English

A Corpus-Based Analysis of the Language of EFL Textbooks used in Secondary Schools in France, Germany and Spain

vorgelegt von

Elen Le Foll

aus Rennes

Osnabrück, März 2022

To my teachers and to all teachers
who believe in their students

Abstract

In secondary English as a Foreign Language (EFL) instructional contexts, textbooks constitute a major and highly influential vector of foreign language input. To date, numerous studies of the language of EFL textbooks have examined textbooks' representations of one or at most a handful of individual linguistic features each. Taken together, these studies provide valuable insights into "the kind of synthetic English" (Römer 2004a: 185) that pupils are exposed to via their school textbooks.

However, the literature review in Chapter 2 makes clear that three crucial aspects have so far been neglected. First, previous research has failed to consider interactions between the frequencies of individual linguistic features. Thus, whilst some influential studies have helped us understand how English learners can be misled by their textbooks into making unidiomatic use of specific linguistic features (e.g., progressive aspect; Römer 2005), only a multivariable approach can paint the full picture as to how "Textbook English" – as a whole – differs from the English that language learners will encounter outside the classroom. Second, prior scholarship has mostly ignored register differences between the various types of texts typically included in school foreign language textbooks. Given that school EFL textbooks frequently feature, for example, extracts from short stories, dialogues, instructions, and exercises on a single double page, this thesis argues that a meaningful analysis of Textbook English requires a register-based approach. Thirdly, previous quantitative corpus-based studies have usually been undertaken at the corpus level, e.g., comparing frequencies across an entire textbook corpus with those from a reference corpus, and have therefore failed to account for the effects of varying textbook proficiency levels or of any potential textbook author, editor, or publisher idiosyncrasies.

The present study therefore sets out to describe the linguistic content of secondary school EFL textbooks and to survey the similarities and most striking differences between the various registers of "Textbook English" and naturally occurring English of situationally similar registers with respect to a wide range of lexico-grammatical features. To this end, the Textbook English Corpus (TEC) was compiled. It comprises nine series of secondary school EFL textbooks (42 textbook volumes) used at lower secondary level in France, Germany and Spain and was manually annotated for six text registers: Conversation, Fiction, Informative texts, Instructional language, Personal correspondence and Poetry & rhyme. In addition, three reference corpora (Spoken BNC2014, Info Teens and Youth Fiction) are used as baselines for comparisons between the language of the TEC and the kind of naturally occurring English that learners can be expected to encounter, engage with, and produce themselves outside the EFL classroom. The compilation of the corpora is outlined in Chapter 3.

Methodologically, a contrastive corpus-based approach is adopted. Chapter 4 reports on a replication of Römer's (2005: Ch. 5–6) study on how the progressive aspect is represented in textbook dialogues compared to naturally occurring conversation, extended to a) include lexical and semantic analyses of the verbs featured in the progressive, and b) cover textbook fiction in addition to conversation. In a second in-depth case study, Chapter 5 explores the lexico-grammatical patterns and semantics of the verb MAKE in textbook conversation and

fiction as compared to the two corresponding target reference corpora. Both chapters point to notable disparities in the use of these features between the Conversation subcorpus of the TEC and the reference conversation corpus. By contrast, Textbook fiction is shown to present these much more authentically.

Chapters 6 and 7 explore the use of multi-dimensional analysis (MDA; Biber 1988; 1995) to model the linguistic specificities of Textbook English. Chapter 6 presents the results of two 'additive' MDAs based on Biber's (1988) model of General Spoken and Written English. The first shows that register accounts for, by far, the greatest proportion of linguistic variation within Textbook English, thus demonstrating that the language of EFL textbooks cannot be adequately modelled without considering register-based variation. By contrast, additional factors such as textbook proficiency level, series, and country of publication/use only play a marginal role in mediating intra-textbook variation. Instructional language is also shown to have very specific linguistic characteristics that set it apart from other textbook registers. The second, contrastive, additive MDA points to a major gap between Textbook Conversation and naturally occurring conversation across all textbook proficiency levels. This is followed by a methodological discussion about the limitations of the traditional MDA framework for the present study. On the basis of these, Chapter 7 proposes a revised MDA framework based on Principal Component Analysis (PCA) and extensive visualisations of the results (inspired by Neumann & Evert 2021). It is trialled with the same data to arrive at a more fine-grained and robust multi-dimensional model of Textbook English. The results of mixed-effects linear regressions modelling the dimension scores of textbook texts and of their corresponding reference corpora largely corroborate those of the two additive MDAs. Together with multi-dimensional visualisations, they are used to explain and illustrate the linguistic specificities of Textbook English.

Chapter 8 provides a summary and general discussion of the results of the four analysis chapters. Lexico-grammatical aspects of Textbook English that substantially diverge from the target reference corpora are highlighted with examples before turning to the study's pedagogical and methodological implications. Suggestions are made as to how teachers, textbook authors and editors could use freely available corpus data and tools to source and modify authentic texts. The case is made for a register approach (see also Rühlemann 2008) to EFL teaching and learning, and implications for teacher training and materials design are discussed. Finally, future research avenues are outlined. These include the triangulation of the present results with learner corpus data to investigate the impact of Textbook English on EFL learners' productive competences.

Zusammenfassung

Im Unterricht für Englisch als Fremdsprache in der Sekundarstufe I spielen Lehrwerke eine zentrale Rolle als Quelle fremdsprachlichen Inputs. Bisher wurde die Sprache der Englischlehrbücher jedoch nur in Bezug auf einzelne linguistische Merkmale untersucht. Diese einschlägigen Untersuchungen liefern grundsätzlich wertvolle Erkenntnisse über „the kind of synthetic English“ (Römer 2004a: 185), mit dem Schüler*innen in Lehrbüchern konfrontiert werden.

Dennoch geht aus dem Literaturüberblick (Kap. 2) hervor, dass die bisherige Forschung drei zentrale Aspekte vernachlässigt. Erstens blieb das Zusammenspiel der Häufigkeiten einzelner linguistischer Merkmale bislang unberücksichtigt. So haben wegweisende Studien gezeigt, wie Englischlernende durch die Schulbuchsprache dazu verleitet werden, bestimmte linguistische Merkmale (wie bspw. die Verlaufsform, vgl. Römer 2005) unidiomatisch zu verwenden. Um einen umfassenden Eindruck zu gewinnen, wie sich das „Schulbuchenglisch“ als Ganzes von dem Englisch unterscheidet, dem Lernende außerhalb des Unterrichts begegnen, ist ein multivariabler Ansatz notwendig. Zweitens wurden die verschiedenen Register, die Lehrwerke für den Fremdsprachenunterricht üblicherweise enthalten, bislang nicht differenziert untersucht. Da solche Lehrbücher auf nur einer Doppelseite oft Textauszüge bspw. aus Dialogen, Kurzgeschichten, Arbeitsanweisungen und Übungen enthalten, wird in dieser Dissertation dargelegt, dass eine aussagekräftige Analyse des Schulbuchenglischen einen Ansatz erfordert, der die Variable Register berücksichtigt. Drittens erfolgten frühere quantitative korpusanalytische Untersuchungen meist auf Gesamtkorpusebene. So wurden bspw. die Frequenzwerte eines gesamten Lehrbuchkorpus mit einem Referenzkorpus verglichen, ohne dabei jahrgangsabhängige Unterschiede im Lernniveau oder durch verschiedene Autor*innen, Herausgeber*innen und Verlage bedingte Besonderheiten der Lehrwerke zu berücksichtigen.

Ziel der vorliegenden Arbeit ist es daher, die sprachlichen Eigenschaften von Lehrwerken für den Englischunterricht in der Sekundarstufe I zu beschreiben. Hierbei soll ein Überblick über die Ähnlichkeiten und die markantesten Unterschiede zwischen verschiedenen Registern des „Schulbuchenglischen“ und den Englischregistern, die in vergleichbaren Situationen außerhalb des Unterrichts verwendet werden, hinsichtlich einer Vielzahl lexiko-grammatischer Merkmale gegeben werden. Hierzu wurde ein Schulbuchkorpus (*Textbook English Corpus*; TEC) bestehend aus neun Lehrwerken (42 Einzelbände) für den Englischunterricht an Sekundarschulen in Frankreich, Deutschland und Spanien erstellt und manuell hinsichtlich sechs Registern annotiert: Konversation, Fiktion, informative Texte, Aufgabenstellungen & Erklärungen, persönliche Korrespondenz und Dichtung & Reim. Zusätzlich dienen drei Referenzkorpora (*Spoken BNC2014*, *Info Teens* und *Youth Fiction*) als Grundlage für den Vergleich zwischen der Sprache des TEC und natürlichem Englisch, dem Lernende außerhalb des Unterrichts begegnen. Der Aufbau der Korpora wird in Kapitel 3 dargestellt.

Methodisch wird hier ein kontrastiver korpusanalytischer Ansatz zugrunde gelegt. Kapitel 4 stellt eine Replikation des Römer (2005: Kap. 5–6) vorgenommen Vergleiches des Gebrauchs der Verlaufsform in Lehrbuchdialogen und in natürlichen Gesprächen dar. Ergänzt wird diese um a) lexikalische und semantische Analysen der in der Verlaufsform verwendeten Verben

und b) die zusätzliche Betrachtung des Registers Fiktion in den Lehrbüchern des TEC. In einer zweiten detaillierten Fallstudie werden in Kapitel 5 die lexiko-grammatischen Konstruktionen sowie die Semantik des Verbs MAKE in den Dialogen und fiktionalen Texten der Lehrbücher im Vergleich zu den beiden entsprechenden Referenzkorpora untersucht. Beide Kapitel belegen erhebliche Unterschiede zwischen der Verwendung dieser Merkmale im Teilkorpus Konversation des TEC und jener im Spoken BNC2014. Dies kontrastiert mit der deutlich authentischeren Wiedergabe des Registers Fiktion in Lehrbuchtexten.

In den Kapiteln 6 und 7 wird die multidimensionale Analyse (MDA, Biber 1988; 1995) für die Modellierung der sprachlichen Besonderheiten des Schulbuchenglischen verwendet. Kapitel 6 enthält dabei die Darstellung der Ergebnisse zweier ‚additiver' multidimensionaler Analysen (MDA) auf Basis von Bibers (1988) Modell des allgemeinen gesprochenen und geschriebenen Englisch. Aus der ersten Analyse ergibt sich, dass der weitaus größte Teil sprachlicher Variation innerhalb des Schulbuchenglischen auf das Register zurückzuführen ist. Dies belegt, dass eine Modellierung der Sprache in Lehrbüchern für den Englischunterricht ohne die Berücksichtigung der Registervariation nicht angemessen ist. Demgegenüber spielen Faktoren wie Unterschiede hinsichtlich Sprachstand, Verlag und Land für Variationen innerhalb eines Schulbuchs eine untergeordnete Rolle. Es zeigt sich ferner, dass sich die spezifischen sprachlichen Merkmale der Aufgabenstellungen und Erklärungen deutlich von anderen Schulbuchregistern unterscheiden. Aus der zweiten, kontrastiven ‚additiven' MDA ergibt sich über alle Kompetenzniveaus der Lehrwerke hinweg ein erheblicher Unterschied zwischen Gesprächen in Schulbüchern und in natürlicher Sprache. Es folgt eine methodische Diskussion der Schwächen des klassischen MDA-Ansatzes für die vorliegende Untersuchung. Auf dieser Grundlage wird in Kapitel 7 ein modifizierter MDA-Ansatz auf Basis einer Hauptkomponentenanalyse und einer umfassenden Visualisierung der Ergebnisse (angelehnt an Neumann & Evert 2021) präsentiert. Umgesetzt wird dies mit denselben Daten, die in Kapitel 6 herangezogen wurden, um zu einem präziseren und robusteren mehrdimensionalen Modell des Schulbuchenglischen zu gelangen. Die Ergebnisse der gemischten Regressionsmodelle der Komponentenwerte von Schulbuchtexten sowie von ihren jeweiligen Referenzkorpora bestätigen diejenigen der beiden additiven MDAs. Zusammen mit den multidimensionalen Visualisierungen werden sie herangezogen, um die sprachlichen Besonderheiten des Schulbuchenglischen zu erklären.

Schließlich erfolgt in Kapitel 8 eine Zusammenfassung sowie eine allgemeine Diskussion der vier Analysekapitel. Lexiko-grammatische Aspekte des Schulbuchenglischen, die deutlich von den jeweiligen Referenzkorpora abweichen, werden hervorgehoben und mit Beispielen illustriert, bevor sich der Fokus einerseits auf fachdidaktische und andererseits auf sprachwissenschaftliche methodische Implikationen richtet. Weiterhin folgen Vorschläge, wie Lehrer*innen, Schulbuchautor*innen und Herausgeber*innen frei verfügbare Korpusdaten und -werkzeuge nutzen können, um an authentische Texte zu gelangen oder diese zu modifizieren. Ein Englischlehr- und -lernansatz, der das Register stärker berücksichtigt (vgl. Rühlemann 2008), wird angeregt und Implikationen für die Lehrer*innenbildung und die Lehrmaterialgestaltung dargestellt. Abschließend werden Perspektiven für zukünftige Forschung aufgezeigt, darunter die Triangulation der vorliegenden Ergebnisse mit Korpora von Lernendendaten zur Untersuchung des Einflusses von Schulbuchenglisch auf die Produktionskompetenz von Lernenden.

Résumé

Dans le domaine de l'enseignement de l'anglais langue étrangère (ALE) dans le secondaire, les manuels scolaires sont un vecteur crucial des contenus linguistiques auxquels sont exposé·e·s les apprenant·e·s. La plupart des études relatives à la langue employée dans les manuels d'ALE n'abordent qu'un seul phénomène linguistique. Dans l'ensemble, ces études offrent des éléments très révélateurs sur le « type d'anglais artificiel » (Römer 2004 : 185) auquel sont exposé·e·s les élèves au travers de leurs manuels.

Cependant, comme l'explique le chapitre 2, les recherches réalisées à ce jour ont délaissé trois aspects clés. Premièrement, elles ont omis d'étudier les interactions entre les différents phénomènes linguistiques. Ainsi, si certaines études faisant autorité ont permis de mieux comprendre pourquoi, sous l'influence de leur manuel scolaire, certain·e·s apprenant·e·s ont tendance à faire un usage peu idiomatique de certains phénomènes linguistiques (l'aspect progressif, p. ex., Römer 2005), seule une approche à variables multiples peut nous permettre de comprendre pleinement les différences entre « l'anglais des manuels » et celui en usage hors des salles de classe. Deuxièmement, jusqu'à présent, la question des différences de registre entre les divers types de textes figurant dans les manuels scolaires a été largement ignorée. Dans la mesure où les manuels d'ALE proposent, p. ex., des extraits d'histoires courtes, des dialogues, des consignes, ainsi que des exercices sur une double page, cette thèse défendra l'idée que toute analyse pertinente de l'anglais des manuels nécessite une prise en compte des registres utilisés. Troisièmement, les études quantitatives sur corpus ont généralement été réalisées au niveau du corpus dans sa globalité, p. ex. en comparant certaines fréquences dans un corpus de manuels avec celles d'un corpus de référence. Ce faisant, elles ont négligé l'impact des différents niveaux de compétence pour lesquels les manuels ont été conçus, ainsi que les choix idiosyncratiques des auteur·rice·s ou directeur·rice·s de publication des manuels.

La présente étude se fixe donc pour objectif de décrire les contenus linguistiques des manuels scolaires d'ALE du secondaire et d'examiner les ressemblances et dissemblances entre les divers registres employés dans l'anglais des manuels et les registres contextuellement similaires en anglais authentique, et cela au regard de toute une série de phénomènes lexicogrammaticaux. A cette fin, nous avons créé un corpus d'anglais des manuels (*Textbook English Corpus* ou TEC) à partir de neuf séries de manuels scolaires d'ALE (42 volumes) utilisés dans les cinq premières années du secondaire en France, en Allemagne et en Espagne. Ce corpus a été annoté manuellement en fonction des registres suivants : Conversation, Fiction, Textes informatifs, Consignes & explications, Correspondance personnelle et Poèmes & rimes. En complément, nous avons utilisé trois corpus de référence (*Spoken BNC2014, Info Teens* et *Youth Fiction*) pour comparer la langue du TEC et l'anglais spontané auquel les élèves sont susceptibles d'être exposé·e·s en dehors de leurs cours d'anglais. La composition des corpus est décrite au chapitre 3.

Du point de vue méthodologique, l'approche adoptée est celle d'une analyse contrastive sur corpus. Le chapitre 4 présente une reprise de l'étude de Römer (2005 : ch. 5–6) sur la manière dont l'aspect progressif s'emploie dans les dialogues des manuels par rapport au discours spontané. L'étude a été étendue sur deux points : a) une analyse lexicale et sémantique des verbes utilisés au progressif et b) la prise en compte du registre de la fiction dans les manuels.

Au travers d'une deuxième étude de cas approfondie, le chapitre 5 examine les phénomènes lexicogrammaticaux et les aspects sémantiques liés au verbe MAKE dans les conversations et les fictions des manuels par rapport aux deux corpus de référence correspondants. Ces deux chapitres mettent en lumière d'importantes disparités quant à l'emploi de ces phénomènes entre le sous-corpus Conversation du TEC et le corpus de conversation de référence. La fiction dans les manuels, quant à elle, en offre une représentation bien plus authentique.

Les chapitres 6 et 7 passent au crible l'utilisation de l'analyse multidimensionnelle (Biber 1988 ; 1995) pour modéliser les spécificités linguistiques de l'anglais des manuels. Le chapitre 6 présente les résultats de deux analyses multidimensionnelles basées sur le modèle de l'anglais général parlé et écrit de Biber (1988). La première montre que le registre est de loin la première source de variation linguistique de l'anglais des manuels, ce qui démontre que la langue employée dans les manuels d'ALE ne peut pas être modélisée correctement sans prendre en compte les variations dues au registre. En revanche, d'autres facteurs, tels que le niveau de compétence pour lequel le manuel a été conçu, la série ou son pays de publication / d'utilisation, ne jouent qu'un rôle minime pour comprendre les variations au sein d'un manuel donné. Nous avons également montré que la langue des consignes et des explications présente des caractéristiques linguistiques très spécifiques, ce qui la différencie des autres registres employés dans les manuels. La seconde analyse multidimensionnelle contrastive met en évidence un sérieux écart entre les conversations dans les manuels et les conversations spontanées, et ce, quel que soit le niveau visé par le manuel. Cette section est suivie d'une réflexion méthodologique sur les limites de l'analyse multidimensionnelle dans la cadre de la présente étude. Sur cette base, le chapitre 7 propose un cadre d'analyse multidimensionnelle révisé reposant sur l'analyse en composantes principales (ACP) et sur des visualisations détaillées des résultats (inspirées par Neumann & Evert 2021). Ce cadre est testé avec les mêmes données dans l'optique de définir un modèle multidimensionnel de l'anglais des manuels plus solide et affiné. Les résultats de modèles mixtes modélisant les scores des différentes dimensions des textes inclus dans les manuels et dans les corpus de référence corroborent largement ceux des deux analyses multidimensionnelles basées sur Biber (1988). Ils sont utilisés en complément de visualisations multidimensionnelles pour expliquer et illustrer les spécificités linguistiques de l'anglais des manuels.

Le dernier chapitre propose un résumé de l'ouvrage ainsi qu'une réflexion générale sur les résultats des quatre chapitres d'analyse. Nous y mettons en évidence, avec des exemples, les aspects lexicogrammaticaux de l'anglais des manuels qui présentent des divergences fondamentales d'avec les corpus de référence avant d'aborder les implications pédagogiques et méthodologiques de cette étude. Nous y faisons des suggestions quant à la manière dont les enseignant·e·s, les auteur·rice·s et directeur·rice·s de publication des manuels pourraient utiliser des données de corpus et des outils de corpus en libre accès pour trouver et adapter des textes plus appropriés en termes de registre. Nous y plaidons en faveur d'une approche basée sur les registres pour l'enseignement et l'apprentissage de l'ALE (cf. Rühlemann 2008) et nous y examinons les implications en matière de formation des enseignant·e·s et de conception des supports. Enfin, nous présentons diverses pistes de recherche futures, telles que la triangulation des présents résultats avec des données de corpus d'apprenant·e·s, pour examiner l'impact de l'anglais des manuels sur les compétences des apprenant·e·s d'ALE en matière de production.

Resumen

En contextos de enseñanza de inglés como lengua extranjera (ILE) en educación secundaria, los libros de texto constituyen un factor fundamental y de extraordinaria influencia en el contacto con el idioma extranjero. Hasta la fecha, numerosos estudios sobre el lenguaje de los libros de texto de ILE han examinado las representaciones de rasgos lingüísticos aislados. En conjunto, estos estudios proporcionan una perspectiva inestimable sobre «the kind of synthetic English [el tipo de inglés sintético]» (Römer 2004: 185) al que los alumnos se ven expuestos a través de sus libros de texto.

Sin embargo, el capítulo 2 demuestra que los estudios anteriores han ignorado tres aspectos cruciales. En primer lugar, no han tenido en cuenta las interacciones entre las frecuencias de aparición de cada uno de los rasgos lingüísticos. De esta forma, si bien algunos estudios influyentes han ayudado a comprender cómo los libros de texto pueden inducir a los alumnos de inglés a usar rasgos lingüísticos específicos de forma poco idiomática (p. ej., el aspecto progresivo; Römer 2005), solo un planteamiento multivariable puede aportar una visión global de cómo el «inglés de libro de texto» difiere del inglés que los alumnos se encontrarán fuera del aula. En segundo lugar, las investigaciones previas han dejado de lado las diferencias de registro entre los distintos tipos de textos que aparecen habitualmente en los libros de texto de idiomas de los centros educativos. Dado que, con frecuencia, los libros de texto de ILE incluyen en una misma página tanto fragmentos de relatos breves como, por ejemplo, diálogos, instrucciones o ejercicios, esta tesis sostiene que un análisis coherente del inglés de libro de texto requiere un enfoque basado en el registro. En tercer lugar, los estudios cuantitativos previos basados en corpus se han realizado a nivel de corpus, p. ej., comparando las frecuencias de todo un corpus de libros de texto con las de todo un corpus de referencia y, por tanto, no han considerado diferencias entre libros de texto, como pueden ser las producidas por factores como el nivel de inglés de los libros, sus autores, redactores o editoriales.

El presente estudio se propone, por tanto, describir el contenido lingüístico de los libros de texto de ILE en educación secundaria y examinar las similitudes y las diferencias más notables entre los diversos registros del «inglés de libro de texto» y el inglés presente de forma natural en situaciones de registro similar para un amplio rango de rasgos léxico-gramaticales. Para tal fin, se ha compilado y anotado manualmente el *Textbook English Corpus* (TEC) compuesto por nueve series de libros de texto de ILE (42 volúmenes) utilizados en el primer ciclo de educación secundaria (CINE nivel 2; UNESCO 2011) en Francia, Alemania y España, que incluye seis registros de texto: conversación, ficción, textos informativos, instrucciones, correspondencia personal, así como poesía y rima. Adicionalmente, se emplean tres corpus (*Spoken BNC2014*, *Info Teens* y *Youth Fiction*) como elementos de referencia para la comparación entre el lenguaje del TEC y el inglés de situaciones reales, es decir, el que cabe esperar que los alumnos encuentren fuera del aula y que ellos mismos producirán. La composición de los corpus se describe en el capítulo 3.

En cuanto a la metodología, se adopta un enfoque contrastivo basado en corpus. El capítulo 4 documenta una réplica del estudio de Römer (2005: Cap. 5–6) sobre la forma de representar el aspecto progresivo en los diálogos de los libros de texto comparada con una conversación

real y ampliada para a) incluir un análisis de los verbos que aparecen en el aspecto progresivo, y b) abarcar la ficción de los libros de texto, además de la conversación. En un segundo estudio de casos exhaustivo, el capítulo 5 explora los patrones léxico-gramaticales y la semántica del verbo MAKE en las conversaciones y la ficción de los libros de texto con respecto a los dos corpus de referencia correspondientes. Ambos capítulos señalan disparidades notables en el uso de estos rasgos entre el subcorpus de Conversación del TEC y el corpus de referencia de conversación. En cambio, muestran que la ficción de los libros de texto los presenta de forma mucho más auténtica.

Los capítulos 6 y 7 exploran el uso del análisis multidimensional (AMD; Biber 1988, 1995) para modelar las características lingüísticas propias del inglés de libro de texto. El capítulo 6 presenta los resultados de dos ADM basados en el modelo de Inglés General Oral y Escrito de Biber (1988). El primero muestra que el registro explica, con gran diferencia, la mayor proporción de la variación lingüística presente en el inglés de libro de texto, demostrando así que el lenguaje de los libros de texto ILE no se puede modelar de forma adecuada sin considerar la variación determinada por el registro. En cambio, otros factores como el nivel de competencia de los libros y el país de publicación o de uso tienen un efecto marginal en la generación de variaciones intra-textuales. Igualmente, el lenguaje de las instrucciones que aparecen en los libros de texto demuestra tener características lingüísticas muy específicas que lo diferencian de otros registros de los libros de texto. El segundo AMD basado en Biber (1988), de índole contrastiva, señala una brecha importante entre la conversación del inglés de libro de texto y la que sucede de forma natural, independientemente del nivel de competencia. A continuación, se aborda una discusión metodológica sobre las limitaciones que supone para este estudio un marco AMD tradicional. Partiendo de estas, el capítulo 7 propone un marco AMD revisado basado en un análisis de componentes principales (ACP), así como en una amplia visualización de los resultados (inspirado en Neumann y Evert 2021). Se utilizan los mismos datos para someter a ensayo dicho AMD revisado, llegando así a dos modelos multidimensionales más robustos del inglés de libro de texto. La aplicación de modelos de regresión lineal con efectos mixtos a los valores de los textos de los libros de texto y sus correspondientes corpus de referencia en las distintas dimensiones arrojan unos resultados que corroboran ampliamente los de los AMD basados en Biber (1988). Estos modelos se emplean, junto con las visualizaciones multidimensionales, para explicar las características lingüísticas propias del inglés de libro de texto.

Por último, el capítulo 8 proporciona un resumen y una discusión general de los resultados de los cuatro capítulos dedicados a los análisis. Antes de pasar a las implicaciones pedagógicas y metodológicas del estudio, se ponen de relieve, mediante ejemplos, diversos aspectos léxico-gramaticales del inglés de libro de texto que divergen de modo sustancial del corpus de referencia. Se presentan distintas sugerencias sobre cómo profesores, autores y editores de libros de texto podrían recurrir a corpus y herramientas de corpus gratuitas para localizar y modificar textos de registro más adecuado. Se exponen asimismo los argumentos en favor de un enfoque de la enseñanza y aprendizaje de ILE basado en el registro (ver también Rühlemann 2008) y se examinan las consecuencias en la formación de los docentes y el diseño de materiales pedagógicos. Finalmente, se esbozan futuras líneas de investigación, tal como la triangulación de los presentes resultados con datos de corpus de alumnos para investigar el efecto del inglés del libro de texto en las destrezas productivas de alumnos ILE.

Acknowledgements

To begin, I would like to thank my *Doktorvater*, Dirk Siepmann, for – what I think is fair to say – was somewhat of a leap of faith in taking me on as his PhD student! For all the ups and downs that this journey entailed, I remain most grateful for this wonderful opportunity. I benefitted tremendously, both academically and personally, from his vast knowledge, mentorship, and trust. I am also very much indebted to Peter Uhrig who took on the role of second PhD supervisor at a particularly difficult time and from whom I learnt a great deal. Thank you both for your support and for all your frank words of advice and wisdom over the past few years.

Before the start of the pandemic, I had the great pleasure to share an office with Anna Fankhauser whose contagiously positive attitude and generosity played a big part in the success of this endeavour. I could not have hoped for a better *Doktorschwester*! Heartfelt thanks also go to Tatjana Winter who, in her capacity as my student research assistant over several semesters, contributed to this project with various annotation tasks and whose interest in the project led to many insightful conversations and the first of hopefully many successful collaborations.

The truth is I have benefited, more or less directly, from the advice and scholarship of far too many people to mention them all here. Yet, at the risk of forgetting someone, I would like to extend special thanks to: Alexander Bergs, Alexandra Elbakyan, Andrea Nini, Carolyn Blume, Dieter Mindt, Douglas Biber, Katharina Delius, Larissa Goulart, Lena Heine, Luke Tudge, Martin Schweinberger, Muhammad Shakir, Natalia Levshina, Stefan Gries, Stephanie Evert, Susanne Flach and Ute Römer. In addition, and with some overlap, I would like to thank the conference attendees of the `Anglistentag`, `DGFF`, `ICAME`, `IVACS` and `TaLC` conferences for their kind welcome into the academic world and generous feedback on my various presentations on this and other research projects.

Years of working remotely can be alienating and, at times, rather demoralising. As such, I am particularly grateful to the `#AcademicTwitter` community – `#AcademicChatter` `#corpuslinguistics` `#TEFL` `#Rstats` and many more – that made `#phdlife` during a global pandemic considerably more enjoyable whilst giving me the opportunity to extend my academic network and broaden my intellectual horizon during my home-office tea breaks in ways that might not have been possible in other circumstances.

I am eternally grateful to my "proofreading squad"[1]: Anja Höing, Anna Fankhauser, Barbara Oberhofer, Birgit Kohn, Charlotte Hahn, Janna Gerdes, Lisa Scheiwe, Marie-Christine Benen, Nathan Dykes, Muhammad Shakir and Susanne Dyka. In addition, it would not have been possible to include translations of the abstract of the thesis in French, Spanish and German without the help of generous friends and colleagues. Un grand merci à Fiona Scuiller, Gaëtanelle Gilquin et Naomi Truan. Muchas gracias a Verónica Lasarte Prieto y Pascual Pérez-Paredes. Und herzlichen Dank an Katharina Hauptmann und Anna Fankhauser. It goes without saying that all remaining errors are mine.

This project has relied on a host of open-source software. As such, I am indebted to the entire Open Source community. In particular, I am grateful for the work of the good people of `RStudio` and `Zotero` and the developers of the countless `R` packages and `python` libraries that I used as part of this project. This PhD would also not have been possible without the contributions of the selfless heroes of `Stack Overflow` & the `R-Ladies` community slack: I hope to be able to pay back my debt in the years to come by helping others take their first scientific coding steps.

Last, but certainly not least, Brompton and his technician have been the most wonderful companions on this long and sometimes arduous PhD journey: thank you for seeing me through. To many more adventures!

[1] With thanks to Nathan Dykes for coining this term.

Table of Contents

1 INTRODUCTION 1

1.1 Aims, scope and methodology 1
1.2 English as a foreign language at secondary school level 3
1.3 Authenticity in EFL teaching 6
1.4 Usage-based theories to L2 learning and teaching 8
1.5 Input and frequency 10
1.6 Input in lower secondary school EFL contexts 11
1.7 Textbooks in the EFL classroom 12
1.8 Corpus linguistics and foreign language education 18
1.9 Outline of the thesis 20

2 LITERATURE REVIEW 23

2.1 Part One: Methodological review 25
2.1.1 Intra-textbook approaches 25
2.1.1.1 Checklist approach to textbook evaluation 25
2.1.1.2 Page-by-page intra-textbook analysis 26
2.1.1.3 Corpus-based intra-textbook analysis 26
2.1.2 Comparative approaches 28
2.1.2.1 Word-frequency lists approaches 29
2.1.2.2 NLP methods 31
2.1.2.3 Corpus-based comparisons of 'real-life' language to textbook language 33
2.1.2.4 Comparing textbook language to 'real-life' language 35
2.1.2.5 Elicitation approaches 37
2.1.2.6 Adding learner corpora to the equation 38
2.1.2.7 Textbook language as learner target language 40
2.1.3 Evaluating the impact of textbook language 42

2.2 Part Two: Key findings of Textbook English studies 44
2.2.1 Lexis 44
2.2.1.1 Individual words 45
2.2.1.2 Multi-word units 47
2.2.2 Tense and aspect 50
2.2.2.1 Future constructions 50
2.2.2.2 The present perfect 50
2.2.2.3 The progressive 51
2.2.2.4 Modals 52
2.2.2.5 Conditionals 52
2.2.2.6 Reported speech 54
2.2.3 Pragmatics 55
2.2.4 Spoken grammar 57

2.3 Conclusion 59

3 RESEARCH AIMS AND DATA 62

3.1 Insights from the literature review 62
3.2 The present thesis 65
3.2.1 Research aims and questions 65
3.2.2 Open Science statement 66

3.3 Corpus data 67
3.3.1 The Textbook English Corpus (TEC) 68
3.3.1.1 Selection of textbooks 68
3.3.1.2 Textbook processing 74
3.3.1.3 Corpus mark-up: headers 74
3.3.1.4 Register annotation 75
3.3.2 The reference corpora 82
3.3.2.1 The Spoken BNC2014 83
On the use of L1 norms in English language teaching 84
Processing of the Spoken BNC2014 88
3.3.2.2 The Youth Fiction corpus 89
3.3.2.3 The Informative Texts for Teens Corpus (Info Teens) 91

4 EXPLORING THE PROGRESSIVE IN TEXTBOOK ENGLISH .. 95

4.1 Rationale and aims 95
4.1.1 EFL learners' use of the progressive 95
4.1.2 The progressive in EFL textbooks 97
4.1.3 Aims and research questions 98

4.2 Methodology 101
4.2.1 Extraction of progressive concordance lines 101
4.2.2 Annotation 103
4.2.2.1 Automatic pre-annotation 103
4.2.2.2 Semantic domains of progressive verbs 104
4.2.2.3 Manual identification of non-progressives 106
4.2.2.4 Morphosyntactic features 106
4.2.2.5 Functions of the progressive 106
4.2.2.6 Validation of the annotation scheme 108
4.2.2.7 Frequency of the progressive 108
4.2.3 Collostructional analysis 111
4.2.3.1 Co-varying Collexeme Analyses: Verb/Semantic domain + Corpus 113
4.2.3.2 Comparative Distinctive Collexeme Analysis 114
4.2.4 Correspondence Analysis 120

4.3 Results and discussion 121
4.3.1 Frequency of progressives 121
4.3.2 Morphosyntactic features 121
4.3.2.1 Tense, aspect and modality 122
4.3.2.2 Contractions 125
4.3.2.3 Negation 127
4.3.2.4 Questions 128
4.3.2.5 Voice 129
4.3.3 Functions of the progressive 131

4.3.3.1 Time reference ... 131
4.3.3.2 Time reference, tense and modality ... 135
4.3.3.3 (Non-)continuousness ... 137
4.3.3.4 Repeatedness ... 140
4.3.3.5 Additional functions of the progressives ... 141
4.3.4 Lexical and semantic aspects ... 142
4.3.4.1 Co-varying Collexeme Analyses of 'Verb lemma' + 'Variety' ... 142
4.3.4.2 Co-varying Collexeme Analyses of 'Semantic domain' + 'Variety' ... 147
4.3.4.3 Correspondence Analysis (CA) of 'Semantic domain' + 'Variety' + 'Register' ... 150
4.3.4.4 Distinctive Collexeme Analyses ... 152

4.4 Conclusion ... 161

5 MAKING SENSE OF MAKE IN TEXTBOOK ENGLISH ... 165

5.1 Rationale and aims ... 165
5.1.1 MAKE in Learner English ... 167
5.1.2 MAKE in Textbook English ... 168

5.2 Methodology ... 169
5.3 Results and discussion ... 171
5.3.1 MAKE in the Textbook English Corpus (TEC) ... 171
5.3.2 Semantics of MAKE in Textbook Conversation and Fiction ... 173
5.3.3 MAKE in the prototypical 'produce' sense ... 176
5.3.4 MAKE as a delexical verb ... 181
5.3.5 Phrasal verbs with MAKE ... 185
5.3.6 Causative MAKE ... 188

5.4 Conclusions ... 192

6 A MULTI-DIMENSIONAL DESCRIPTION OF TEXTBOOK ENGLISH ... 195

6.1 Introduction ... 195
6.1.1 Multi-feature/multi-dimensional analysis (MDA) ... 196
6.1.2 MDA and textbook language ... 207
6.1.2.1 Exploring Textbook English using Additive MDA ... 207
6.1.2.2 Exploring Textbook English by conducting a full MDA ... 209

6.2 Method ... 210
6.2.1 Choosing a baseline MDA study ... 210
6.2.2 Defining text units in the TEC ... 211
6.2.3 Tagging and counting features ... 213
6.2.4 Computing the mean dimension scores for the new registers ... 214
6.2.5 Computing dimension scores for additional reference corpora ... 214
6.2.6 Comparing dimension scores ... 215

6.3 Results ... 219
6.3.1 Intra-textbook linguistic variation ... 220

6.3.1.1 Intra-textbook variation on Biber's Dimension 1 ... 220
6.3.1.2 Intra-textbook variation on Biber's Dimension 2 ... 228
6.3.1.3 Intra-textbook variation on Biber's Dimension 3 ... 233
6.3.1.4 Intra-textbook variation on Biber's Dimension 4 ... 239
6.3.1.5 Intra-textbook variation on Biber's Dimension 5 ... 241
6.3.1.6 Intra-textbook variation on Biber's Dimension 6 ... 244
6.3.2 Comparative additive MDA ... 248
6.3.2.1 The specificities of Textbook English on Biber's (1988) Dimension 1 ... 248
6.3.2.2 The specificities of Textbook English on Biber's (1988) Dimension 2 ... 257
6.3.2.3 The specificities of Textbook English on Biber's (1988) Dimension 3 ... 260

6.4 Discussion ... 262
6.5 Limitations ... 266

7 TOWARDS A NEW MULTI-DIMENSIONAL UNDERSTANDING OF TEXTBOOK ENGLISH ... 269

7.1 Introduction ... 269
7.2 Methodology: the MDA framework revised ... 269
7.2.1 Design bias in the selection of text samples ... 271
7.2.2 Design bias in the choice and operationalisation of features ... 272
7.2.3 Uncertainty over the reliability of the feature counts ... 276
7.2.4 Confusion between covariation of features due to situational variation and covariation due to grammatical structure ... 281
7.2.5 Lack of transparency in the quantitative patterns captured by factor analysis ... 284
7.2.6 Degradation of correlations ... 288
7.2.7 Arbitrary thresholds in the computation of dimension scores ... 295
7.2.8 Misleading visualisation of results ... 298
7.2.9 Difficulties in establishing the statistical significance and robustness of the results ... 300
7.2.10 Issues related to reproducibility and replicability ... 301

7.3 Results ... 302
7.3.1 Linguistic variation within the TEC ... 302
7.3.1.1 Variation along Dimension 1: 'Overt instructions and explanations' ... 312
7.3.1.2 Variation on Dimension 2: 'Involved vs. Informational Production' ... 317
7.3.1.3 Variation on Dimension 3: 'Present/factual vs. Past/speculative' ... 323
7.3.1.4 Variation on Dimension 4: 'Clausal complexity' ... 324
7.3.2 PCA of Textbook English vs. 'real-life' English ... 327
7.3.2.1 Textbook Conversation vs. the Spoken BNC2014 ... 341
7.3.2.2 Textbook Fiction vs. the Youth Fiction corpus ... 347
7.3.2.3 Textbook Informative vs. Info Teens ... 350

7.4 Discussion and limitations ... 353

8 GENERAL DISCUSSION AND IMPLICATIONS ... 361

8.1 Summary and conclusions ... 361
8.2 Pedagogical implications and recommendations ... 370
8.2.1 Improving representations of conversational English ... 373
8.2.2 Improving representations of informative texts ... 381

8.2.3 Towards a register approach to teaching EFL 388
8.2.4 Implications for teacher education 391
8.2.5 Implications for materials design 396

8.3 Limitations and methodological implications 400
8.4 Future research avenues 403
8.5 Conclusion 404

REFERENCES 406

APPENDICES 452

List of Figures

Fig. 1: Proportion of tokens in the three "national" subcorpora the TEC (as displayed by Sketch Engine; Kilgarriff et al. 2014) 72
Fig. 2: Proportion of tokens in each register category on the TEC 79
Fig. 3: Composition of the web corpus originally retrieved by Sketch Engine 92
Fig. 4: Distribution of progressive forms in Textbook Conversation and the Spoken BNC2014 122
Fig. 5: Association plot comparing the progressive form distributions in Textbook Conversation and the Spoken BNC2014 sample 123
Fig. 6: Progressive form distributions in the Spoken BNC2014 sample and the five different Textbook levels 124
Fig. 7: Association plot comparing the tense form distributions in Textbook Fiction and the Youth Fiction sample 125
Fig. 8: Proportion of contracted progressives by progressive form in Textbook Conversation as compared to the Spoken BNC2014 sample 126
Fig. 9: Time period reference for the progressives in Textbook Conversation and the Spoken BNC2014 sample 132
Fig. 10: Association plot of the distribution of time references of progressives in Textbook Conversation and the Spoken BNC2014 sample 133
Fig. 11: Distribution of time references of progressives in the Spoken BNC2014 sample and in Textbook Conversation for each textbook level 134
Fig. 12: Association plot of progressive time reference between Level C-E Textbook Conversation and the Spoken BNC2014 sample 135
Fig. 13: Mosaic plot displaying the cross-tabulation of forms and time references of progressives in Textbook Conversation and the Spoken BNC2014 sample (progressives with 'unclear' time references were removed for this analysis). 136
Fig. 14: Distribution of the semantic categories of verbs in the progressives across the four (sub)corpora (excluding polysemous verbs). 148
Fig. 15: Simple binary correspondence analysis of 'Semantic domain' and 'Corpus' 151
Fig. 16: Comparison of the standardised CL values (scaled and centred G^2) returned by the two progressive vs. non-progressive Distinctive Collexeme Analyses (DCAs). Positive values represent attraction to the progressive, negative to the non-progressive construction. 154
Fig. 17: Comparison of the standardised CL values (scaled and centred G^2) returned by the two progressive vs. non-progressive Distinctive Collexeme Analyses (DCAs) 159
Fig. 18: Relative frequencies of the verb lemma MAKE in Textbook and Reference subcorpora 173
Fig. 19: Comparison of the distribution of MAKE meanings in Textbook Conversation and the Spoken BNC2014 sample 173
Fig. 20: Comparison of the distribution of MAKE meanings in Textbook Fiction and Youth Fiction 175
Fig. 21: Distribution of the semantic fields attributed to the noun collocates of MAKE in the (Ai) 'produce' sense in Textbook Conversation and the Spoken BNC2014 sample 176
Fig. 22: Frequent collocate lemmas of MAKE in the (Ai) 'produce' sense in Textbook Conversation and Spoken BNC2014 sample 177
Fig. 23: Differences in semantic fields attributed to collocates of MAKE in the (Ai) 'produce' sense. 180
Fig. 24: Most frequent collocates of delexical MAKE in Textbook Conversation and the Spoken BNC2014 182
Fig. 25: Most frequent collocates of delexical MAKE in Textbook Fiction (left) and the Youth Fiction sample (right) 183

Fig. 26 'Speech/communication' delexical MAKE collocations in Textbook Conversation and Textbook Fiction (red) as compared to the reference Spoken BNC2014 and Youth Fiction samples (beige). 184
Fig. 27: The most frequent collexemes of the causative MAKE constructions of Textbook Conversation and the Spoken BNC2014 sample 190
Fig. 28: Correlation matrix of the normalised counts in Table 39 199
Fig. 29: Mean scores of general spoken and written registers of English on Biber's (1988) Dimension 1 (as summarised in Biber & Conrad 2019: 292) 206
Fig. 30: Matrix of correlations among the estimated random effects of the maximal model: lmer(Dim2 ~ Register + Level + Level*Register + (Register|Series)) 217
Fig. 31: Distribution of texts of the TEC on Biber's (1988) Dimension 1 221
Fig. 32: Observed (grey circles) and predicted (red triangle) Dimension 1 scores across textbook register, proficiency level and textbook series. Predicted values as computed by the model summarised in Table 66. (Le Foll 2021. Zenodo. Retrieved on 7 May 2021. http://doi.org/10.5281/zenodo.4732323) 224
Fig. 33: Distribution of the texts of the TEC on Biber's Dimension 2 (excluding one outlier text from the Poetry register) 229
Fig. 34: Visualisation of the effects of Register*Level interactions on the estimated Dimension 2 values of the TEC texts (as calculated with the model summarised in Table 49) 232
Fig. 35: Distribution of the texts of the TEC on Biber's (1988) Dimension 3 234
Fig. 36: Estimated scores on Biber's (1988) Dimension 3 subdivided by register and textbook series 238
Fig. 37: Distribution of the texts of the TEC on Biber's (1988) Dimension 4 (with one outlier from the Informative register removed) 239
Fig. 38: Estimated mean Dimension 4 scores subdivided per textbook register and proficiency level 241
Fig. 39: Distribution of the texts of the TEC on Biber's (1988) Dimension 5 242
Fig. 40: Distribution of the texts of the TEC on Biber's (1988) Dimension 6 244
Fig. 41: Estimated scores on Biber's (1988) Dimension 6 across levels C to E textbooks 247
Fig. 42: Comparison of the Conversation, Fiction and Informative texts from the TEC with the three corresponding target language reference corpora on Biber's (1988) Dimension 1 (as calculated by the MAT, see below) 248
Fig. 43: Comparison of scores on Biber's (1988) Dimension 1 of the Spoken BNC2014 and Textbook Conversation, recalculated so as to exclude the features that rely on punctuation marks for their operationalisation 250
Fig. 44: Estimated varying effects of textbook series on modified Dimension 1 scores for Textbook Conversation (simulated using the R function merTools::REsim; Knowles & Frederick 2020) 254
Fig. 45: Informative textbook text with a low Dimension 1 score <TEC: Join the Team 4^{e}> 255
Fig. 46: Comparison of the Conversation, Fiction and Informative texts from the TEC with the three corresponding target language reference corpora on Biber's (1988) Dimension 2 257
Fig. 47: Normalised frequencies of features loading on Biber's (1988) Dimension 2 for Textbook Fiction (subdivided by proficiency level A to E) and Youth Fiction 260
Fig. 48: Comparison of the Conversation, Fiction and Informative texts from the TEC with the three corresponding target language reference corpora on Biber's (1988) Dimension 3 261
Fig. 49: Per-feature accuracy measures of the MFTE on the TEC, the Spoken BNC2014 and data comparable to the Youth Fiction and Info Teens 279
Fig. 50: The ten most frequent word forms occurring immediately after *if* in the Spoken BNC2014 (as counted and displayed by Sketch Engine, see https://ske.li/nu8 for full results) 282
Fig. 51: Distribution of normalised frequencies of occurrence of five features across the TEC (histograms) and visualisations of their correlation (scatterplots) 290

Fig. 52: Distribution of signed log transformed standardised normalised frequencies of occurrence of five features across the TEC (histograms) and visualisations of their correlation (scatterplots) 292
Fig. 53 Correlation matrix of the signed log standardised relative frequencies of the features analysed in the second PCA-based MDA (see 7.3.2) 295
Fig. 54: Dimension 1 mean scores for disciplines (left) and genre families (right) from Gardner et al. 299
Fig. 55: Scree plot of the eigenvalues of the principal components (PCs) for the TEC data 303
Fig. 56: Snapshots from the 3-D visualisation of the first three dimensions of the multi-dimensional model of intra-textbook variation 304
Fig. 57: Scatterplot matrix of combinations of the first six dimensions of the model of intra-textbook variation (the number before the comma on each axis label shows which principal component is plotted on that axis; this is followed by the percentage of the total variance explained by that particular component) 305
Fig. 58: Projection of the texts of the TEC on the first and second dimensions of the model of intra-textbook variation 307
Fig. 59: Graph of the features with the strongest contributions to the first and second dimensions of the model of intra-textbook variation 307
Fig. 60: Projection of the texts of the TEC on the third and fourth dimensions of the model of intra-textbook variation 309
Fig. 61: Graph of the features with the strongest contributions to the third and fourth dimensions of the model of intra-textbook variation 309
Fig. 62: Projection of the texts of TEC on third and fourth dimensions with colours and ellipses indicating the proficiency level of the textbooks (as opposed to register as in Fig. 60) 311
Fig. 63: Projection of the texts of the TEC on the fifth and sixth dimensions of the model of intra-textbook variation 311
Fig. 64: Coefficient estimates and 95% confidence intervals of the fixed effects in the model: lmer(PC1 ~ Register + (1|Series)) (the intercept corresponds to the reference level: Register [Conversation]) 315
Fig. 65: Coefficient estimates and 95% confidence intervals of the fixed effects in the model: lmer(PC2 ~ Register + Level + Register*Level + (1|Series)). The reference levels are Register [Conversation] and Level [A] 319
Fig. 66: Estimated PC2 scores across each register and the five textbook proficiency levels 321
Fig. 67: Coefficient estimates and 95% confidence intervals of the fixed effects in the model: lmer(PC3 ~ Register + Level + Register*Level + (1|Series). The intercept corresponds to the reference levels Register [Conversation] and Level [A]. 323
Fig. 68: Estimated PC3 scores across each register and the five textbook proficiency levels 324
Fig. 69: Coefficient estimates and 95% confidence intervals of the fixed effects in the model: lmer(PC4 ~ Register + Level + Register*Level + (1|Series)). The reference levels are Register [Conversation] and Level [A] 326
Fig. 70: Estimated PC4 scores across each register and the five textbook proficiency levels 327
Fig. 71: Scree plot of the eigenvalues of the PCs for the Textbook English vs. 'real-life' English PCA 328
Fig. 72: Snapshots from the 3-D representation of texts along PC1–PC3 329
Fig. 73: Scatterplot matrix of combinations of the four dimensions of the model of Textbook English vs. 'real-life' English (the number before the comma on each axis label shows which principal component is plotted on that axis; this is followed by the percentage of the total variance explained by that particular component) 330
Fig. 74: Projection of the texts of the three subcorpora of the TEC and the reference corpora on PC1 and PC2 332
Fig. 75: Graph of the features with the strongest contributions to the first and second dimensions 332

Fig. 76: Projection of the texts of the three subcorpora of the TEC and the reference corpora on PC3 and PC4 335
Fig. 77: Projection of the texts of the three subcorpora of the TEC and the reference corpora on PC3 and PC4 with ellipses representing the five textbook proficiency levels vs. the reference corpora 335
Fig. 78: Predicted PC3 scores of the texts of the TEC and the reference corpora 337
Fig. 79: Features that make the most important contributions to the third and fourth dimensions 338
Fig. 80: Predicted PC4 scores of the texts of the TEC and the reference corpora 340
Fig. 81: Predicted PC1 scores of the texts of the TEC and the reference corpora 342
Fig. 82: Predicted PC2 scores of the texts of the TEC and the reference corpora 350
Fig. 83: Normalised counts of selected features with salient loadings on PC1 in the Textbook Informative subcorpus (Levels A to E) and the reference Info Teens corpus (Ref.) 352
Fig. 84: Proportion of tokens in each register category of the French subcorpus of the TEC (as displayed by Sketch Engine; Kilgarriff et al. 2014) 356
Fig. 85: Cumulative word counts of the Conversation, Fiction, Instructional, Informative, Personal correspondence and Poetry texts for each of the 42 textbook volumes of the TEC 357
Fig. 86: Projection of texts on PC1 and PC2 from a random 2/3 split-data analysis of the three subcorpora of the TEC and the three reference corpora 365
Fig. 87: Projection of texts on PC3 and PC4 from a random 2/3 split-data analysis of the three subcorpora of the TEC and the three reference corpora 365
Fig. 88: Exercise featured below (254) 383
Fig. 89: Grammar box featured below (254) 383
Fig. 90: Word frequency analysis with english-corpora.org (on the basis of COCA data) of excerpt (255) 384
Fig. 91: Part of the 'word sketch' page of the word *lush* as generated on english-corpora.org/coca. 385
Fig. 92: Word frequency analysis with english-corpora.org (on the basis of COCA data) of excerpt (254) 386
Fig. 93: Part of the "word sketch" page of the word *moreover* on english-corpora.org/coca..... 386

List of Tables

Table 1: Distribution of the articles *a, an* and *the* in the five Malaysian textbooks examined by Mukundan et al. 2012 (as reported in Table 1, p. 69) and in the BNC1994 (as calculated by Sketch Engine on 13.10.2018) in raw numbers and as a % of articles 27
Table 2: Comparison of the order of the most frequent prepositions in the BNC1994 and three Malaysian ESL textbooks 31
Table 3: The levels of the Textbook English Corpus (TEC) 69
Table 4: Most widely used lower secondary school textbook series (and publisher in brackets) according to the informal market surveys conducted with teachers, bookshop assistants and publishers in France, Germany and Spain 70
Table 5: Composition of the Textbook English Corpus (TEC) (the full bibliographic metadata is available on doi.org/10.5281/zenodo.4922819) 73
Table 6: Number of words per textbook register categories in the TEC (as calculated by Sketch Engine Kilgarriff et al. 2014) 79
Table 7: Summary of the regular expressions (regex) used to process the Spoken BNC2014 (see Online Appendix 3.3 for full script) 89
Table 8: Composition of the Informative Texts for Teen Corpus 93
Table 9: Core and additional function categories for the progressive (all examples from Römer 2006b: Appendix) 107
Table 10: Example contingency table for the verb SAY in the first CovCA conducted (see 4.2.3.1) 114
Table 11: Last three entries of the count table for the conversation register before normalisation .. 116
Table 12: Contingency table for SAY in Textbook Conversation (the figures in bold were directly counted from the subcorpus) 116
Table 13: Contingency table for SAY in the Spoken BNC2014 sample (the figures in bold have been directly counted, underlined approximated from the corpus) 117
Table 14: Examples of lemmas whose non-progressive FVP counts were negative in the Spoken BNC2014 118
Table 15: Example lemma frequency counts for the Spoken BNC2014 after having adjusted for sampling/tagging errors 119
Table 16: Contingency table for SAY in the Spoken BNC2014 post-adjustments 119
Table 17: Proportion (and raw numbers) of contracted forms of the auxiliary BE in progressive constructions 126
Table 18: Proportions of contracted and non-contracted progressives in Textbook Fiction and the Youth Fiction corpus 127
Table 19: Proportion of negated progressives in Textbook Conversation and the Spoken BNC2014 sample 127
Table 20: Proportion of negated progressives in Textbook Fiction texts and the Youth Fiction sample 127
Table 21: Voice of progressive constructions 130
Table 22: The continuousness function of progressives in the two textbook registers and reference corpora 138
Table 23: Most frequent lemmas occurring in the progressive with a 'non-continuous' function in Textbook Conversation and the Spoken BNC2014 (total occurrences in 'non-continuous' function) (as a percentage of all progressive occurrences) 139
Table 24: The repeatedness function of progressives in the two textbook registers and reference corpora as a percentage of all progressives (and in absolute figures) 140
Table 25: Percentages of progressives within each (sub)corpus coded for additional functions (raw figures are given in brackets) 141

Table 26: Top 20 Co-varying Collexemes for 'Verb' + 'Variety' (Spoken BNC2014 vs. Textbook Conversation) 143
Table 27: Top 10 Co-varying Collexemes for '*Verb' + 'Variety'* (Textbook Fiction vs. Youth Fiction sampled) 146
Table 28: Results of the CovCA 'Semantic domain' + 'Variety' (Textbook Conversation vs. reference Spoken BNC2014 sample) 149
Table 29: Eight most frequent 'communication' verbs in the Spoken BNC2014 progressive concordance sample ($n = 370$) 149
Table 30: Results of the CovCA of *'Semantic domains' + 'Fiction Varieties'* (Textbook Fiction vs. Youth Fiction sample) 150
Table 31: Most significantly distinctive collexemes associated with the progressive (positive G^2 values) and the non-progressive constructions (negative G^2 values) in Textbook Conversation and the Spoken BNC2014 sample 153
Table 32: Top 30 verb lemmas with the greatest absolute difference in attraction/repulsion CL scores to the progressive in Textbook Fiction and the Youth Fiction Corpus (negative values in the second and/or third columns represent repulsion of the progressive, in other words, attraction to non-progressive constructions) 158
Table 33: Semantic categories of MAKE according to the VDE (Herbst, Heath & Roe 2013) illustrated with sentences from the TEC 170
Table 34: Frequency of the verb lemma MAKE in the TEC 172
Table 35: Distribution of the most frequent idioms featuring MAKE in Textbook Conversation and the Spoken BNC2014 sample (absolute frequencies) 174
Table 36: Absolute frequencies of phrasal verbs with MAKE (out of 674 occurrences of MAKE each for the Conversation samples and 392 for the two Fiction samples) 186
Table 37: Causative MAKE constructions in the Textbook Conversation and Spoken BNC2014 sample 189
Table 38: Linguistic features used in Biber's (1988) MDA of general English (as categorised and listed in Conrad & Biber 2013: 18–19) 197
Table 39: Selected normalised feature counts (per 100 words in three texts (see excerpts (179)–(181) below) 199
Table 40: Features with a minimum factor loading of ±0.35 that make up Biber's (1988) seven-factor solution 202
Table 41: The computation of dimension scores on the basis of normalised frequencies 204
Table 42: The computation of dimension scores on the basis of standardised frequencies (z-scores) 204
Table 43: Summary of Biber's six dimensions of English (1988) 205
Table 44: Distribution of Textbook English Corpus (TEC) texts processed in this chapter 213
Table 45: Summary of the model: lmer(Dim1 ~ Register + Level + Level*Register + (1|Series)) ... 222
Table 46: Estimated register means of Dimension 1 scores (averaged across all textbook levels) ... 226
Table 47: Estimated differences between mean Dimension 1 scores for each TEC register pair (averaged across all textbook levels) 226
Table 48: Mean z-scores of the features that load on Biber's (1988) Dimension 1 227
Table 49: Summary of the model: lmer(Dim2 ~ Register + Level + Level*Register + (1|Series)) ... 229
Table 50: Estimated Dimension 2 means and 95% confidence intervals (CI) 233
Table 51: Summary of the model: lmer(Dim3 ~ Register + Level + Level*Register + (1|Series)) ... 234
Table 52: Summary of the model: lmer(Dim4 ~ Register + Level + Level*Register + (1|Series)) (excluding the outlier discussed above) 240
Table 53: Mean Dimension 5 scores for Informative texts grouped by textbook level 243
Table 54: Summary of the model: lmer(Dim5 ~ Register + Level + Level*Register + (1|Series)) for Levels C to E textbook texts only 243
Table 55: Summary of the model: lmer(Dim6 ~ Register + Level + Level*Register + (1|Series)) for levels C to E textbooks only 246

Table 56: Summary of the model: lmer($Dim1_{adjusted}$ ~ 1 + Corpus + Register + Corpus*Register + (Register|Source)) 250
Table 57: Normalised counts for the features loading on Biber's (1988) Dimension 1, except those that rely on punctuation for their operationalisations (features with positive loadings in red and bold, with negative loadings in blue) 251
Table 58: Summary of the model: lmer(Dim2 ~ 1 + Corpus + Register + Corpus*Register + (Register|Source)) 258
Table 59: Estimated Dimension 2 means (degrees-of-freedom method: asymptotic) 258
Table 60: Summary of the model: lmer(Dim3 ~ 1 + Corpus + Register + Corpus*Register + (1|Source)) 261
Table 61: Estimated mean Dimension 3 scores (degrees-of-freedom method: asymptotic) 262
Table 62: Textbook English Corpus (TEC) texts processed in this chapter 272
Table 63: Excerpt of Appendix I: Operationalisation of the 'DO as an auxiliary' feature 275
Table 64: Summary of the terminology used in tagger performance evaluation 277
Table 65: Most frequent tagging error types 280
Table 66: Summary of the model: lmer(PC1 ~ Register + Level + Level*Register + (1|Series)) 312
Table 67: Estimated differences between mean PC1 scores for each TEC register pair (averaged across all textbook levels and series) 314
Table 68: List of feature loadings (eigenvectors) on the four dimensions of the model of intra-textbook variation 315
Table 69: List of feature loadings (eigenvectors) in the Textbook English vs. 'real-life' English PCA-based model 333
Table 70: Summary of the model: lmer(PC3 ~ 1 + Level + Register + Level*Register + (1|Source)) 336
Table 71: Summary of the model: lmer(PC1 ~ 1 + Level + Register + Level*Register + (Register|Source)) 341
Table 72: Summary of the model: lmer(PC2 ~ 1 + Level + Register + Level*Register + (1|Source)) 348

List of Abbreviations

BNC	British National Corpus
CA	Correspondence analysis (see 4.2.4)
CL	Collostructional strength (see 4.2.3)
COCA	Corpus of Contemporary American English
CovCA	Co-varying collexeme analysis (see 4.2.3.1)
DCA	Distinctive collexeme analysis (see 4.2.3.2)
EAP	English for academic purposes
EFA	Exploratory factor analysis (see 7.2.5)
EFL	English as a foreign language (see 1.1)
EGP	English for general purposes
ELF	English as a lingua franca (see 3.3.2.1)
ENL	English as a native language (see 1.1)
ESL	English as a second language (see 1.1)
ESP	English for specific purposes
FA	Factor analysis (see 7.2.5)
FPV	Finite verb phrase (see 4.2.2.7)
Info Teens	Informative Texts for Teens Corpus (see 3.3.2.3)
L1	First language(s) (see 1.1)
L2	Second or more (foreign) language (see 1.1)
MAD	Median absolute deviation
MFTE	Multi-feature tagger of English (see 7.2.2)
PCA	Principal component analysis (see 7.2.5)
pmw	per million words
POS	Part of speech
SD	Standard deviation
TEC	Textbook English Corpus (see 3.3.1)
VDE	Valency Dictionary of English (Herbst, Heath & Roe 2013)
Youth Fiction	Youth Fiction Corpus (see 3.3.2.2)

Note that the full list of the linguistic features tagged by the MFTE and their abbreviations (see Chapter 7) can be found in Appendix I.

1 Introduction

[T]he reason that coursebooks are so often in the line of fire is that they do to a large extent dominate and determine so many aspects of a teacher's day-to-day professional life.
❖ Scott Thornbury (2012)

Asked "Where is Brian?", French nationals of a certain generation will immediately reply: "Brian is in the kitchen". Those with a particularly good memory may follow up with: "Where is Jenny, the sister of Brian?" – and, to those in the know, the correct answer is: "Jenny is in the bathroom".[2] Clearly, there is no need for any in-depth linguistic analysis to conclude that this interaction is highly unlikely to have ever taken place in a real English-speaking family home. To most teachers and learners, it will be evident that it is the result of a none too inspired attempt to model WH-question forms in a beginner textbook dialogue aimed at learners of English as a Foreign Language (EFL). Together with dull gap-fill exercises and photos of out-of-date technology, for many adults, the very mention of the word *textbook* evokes vivid memories of such artificially sounding, contrived and sometimes even non-sensical dialogues.

This raises the question of the status and nature of textbook language as a specific 'variety' of language, which is the subject of the present study. It focuses on contemporary EFL textbooks in use in European secondary schools. Situated at the interface between linguistics and foreign language education, it examines the linguistic content of these textbooks and seeks empirical answers to the questions: What kind of English do school EFL textbooks portray? And how far removed is this variety of English from the kind of English that learners can be expected to encounter outside the EFL classroom?

1.1 Aims, scope and methodology

The above questions are critical because, as many adults' lingering memories of school foreign language lessons testify (see also, e.g., Freudenstein 2002: 55), textbooks play an absolutely central role in classroom-based foreign language learning. In the following, we will see that the dominance of textbooks in EFL school contexts persists to this day. According to Thornbury (2012 in a comment response to Chong 2012: n.p.), they "(more often [than] not) instantiate the curriculum, provide the texts, and - to a large extent - guide the methodology".

[2] Dialogue from *Speak English 6e série verte* (Benhamou & Dominique 1977: 167). It was made popular by stand-up comedian Gad Elmaleh.

In lower secondary EFL instructional contexts, in particular, textbooks constitute a major vector of foreign language input. Yet, numerous studies have shown that "considerable mismatches between naturally occurring English and the English that is put forward as a model in pedagogical descriptions" (Römer 2006a: 125–126) exist. These mismatches have been observed and sometimes extensively described in textbooks' representations of numerous language features ranging from the use of individual words and phraseological patterns (e.g., Conrad 2004 on the preposition *though*; Gouverneur 2008a on the high-frequency verbs *make* and *take*), to tenses and aspects (e.g., Barbieri & Eckhardt 2007 on reported speech; Römer 2005 on the progressive). More rarely, textbook language studies have also ventured into the study of spoken grammar (e.g., Gilmore 2004) and pragmatics (e.g., Hyland 1994 on hedging in ESP/EAP textbooks). As will be shown in Chapter 2, previous EFL textbook studies have tended to focus on one or at most a handful of individual linguistic features. Taken together, they provide valuable insights into "the kind of synthetic English" (Römer 2004a: 185) that pupils are exposed to via their textbooks; yet, what is missing is a more comprehensive, broader understanding of what constitutes "Textbook English" from a linguistic point of view. In particular, although corpus-based[3] textbook analysis can be traced back to the pioneering work of Dieter Mindt in the 1980s, the language of secondary school EFL textbooks (as opposed to that of general adult EFL or English for Specific Purposes [ESP] coursebooks) remains an understudied area.

The present study therefore sets out to describe the linguistic content of secondary school EFL textbooks and to survey the similarities and most striking differences between "Textbook English" and "naturally occurring English" with respect to a wide range of lexico-grammatical features. To this end, a corpus of nine series of secondary school EFL textbooks (42 textbook volumes) used at lower secondary level in France, Germany and Spain was compiled (see 3.3.1). In addition, three reference corpora are used as baselines for comparisons between the language input EFL learners are confronted with via their school textbooks and the kind of naturally occurring English that they can be expected to encounter, engage with, and produce themselves on leaving school. Two of these have been built specifically for this project with the aim of representing comparable 'authentic' (for a discussion of this controversial term in ELT, see 1.3) and age-appropriate learner target language.

As hinted at in the title, a bottom-up, corpus-based approach is adopted (e.g., Mindt 1992; 1995a; Biber & Quirk 2012; Biber & Gray 2015; Carter & McCarthy 2006a). Various corpus-linguistic methods are used to analyse the linguistic specificities of Textbook English. A broad range of linguistic features ranging from tenses and

[3] In the present work, the adjectives 'corpus-based' and 'corpus-driven' are used synonymously (see, e.g., Meunier & Reppen 2015: 499 for further information as to how these terms are sometimes distinguished).

aspects to negation and discourse markers are examined in both quantitative and qualitative contrastive analyses. Lexico-grammatical aspects of Textbook English that substantially diverge from the target learner language reference corpora are highlighted and illustrated with direct comparisons of textbook excerpts with comparable texts from the reference corpora.

The rest of this chapter briefly outlines the underlying pedagogically-driven and theoretical motivations for this study beginning with the status of English as a foreign language (EFL)[4] at secondary school level with a focus on continental Europe. This is followed by a summary of some of the controversies around the contentious concept of 'authenticity' in 1.3. An overview of the linguistic and language development theories that motivated the present study is represented in 1.4, followed by a section highlighting the centrality of input in foreign language learning and teaching in 1.5. Section 1.6 turns to language input in lower secondary school EFL contexts in particular, whilst 1.7 focuses on secondary school EFL learners' main source of English input: their textbooks. Finally, Section 1.8 situates the present study and its methodological framework within the growing body of "pedagogically-driven corpus-based research" (Gabrielatos 2006: 1). This introductory chapter concludes with an outline of the rest of the thesis in 1.9.

1.2 English as a foreign language at secondary school level

As the most widely taught foreign language and the *lingua franca* of choice in business and academia, the upmost relevance of English as a Foreign Language (EFL) in the 21st century need not be explained. In mainland Europe, too, English is by far the most widely taught foreign language. At lower secondary school level, defined here as ISCED 2 (UNESCO Institute for Statistics 2012; 2015) (see also 3.3.1.1), almost all students (97.3% according to the latest available figures from 2014) attend English classes (European Commission, EACEA, & Eurydice 2017: 13).

[4] Note that, throughout this thesis, the term 'English as a foreign language' (EFL) is used to refer to learning English in countries and regions where English is not an official or otherwise widely used language (e.g., France, Germany and Spain). For non-native English speakers learning English in countries/regions where English is an official or widely used language, the term 'English as a second language' (ESL) is preferred. Where both learning contexts are meant, English 'L2' is used, regardless of whether English is in fact an individual's second, third, or more non-native language. The author recognises that all of these terms – 'native' vs. 'non-native' or 'foreign' and 'L1' vs. 'L2' – are inherently problematic with regards to their epistemology, operationalisations and underlying assumptions (see, e.g., Birkland et al. 2022; Holliday 2005; Ramjattan 2019). For lack of a better generalisable categorisation system, however, the terms 'English native speaker'/'English L1 user' and 'EFL learner'/'English L2 user' are used throughout the present thesis as an imperfect means to differentiate between two typically very different language acquisition contexts in full recognition that such a dichotomisation represents a vast over-simplification of what are frequently much more complex language biographies and learning experiences.

In Germany, English is a mandatory subject during compulsory secondary education in nine out of the sixteen *Bundesländer* (European Commission, EACEA, & Eurydice 2017: 44). As of 2013, 97.8% of students were learning English in the *Sekundarstufe I*. Similarly high rates were recorded in France (98.6%) and Spain (100%) in 2016 (European Commission, EACEA, & Eurydice 2017: 164). Across Europe, a clear upward trend in the proportion of students learning English in compulsory school education can be observed for the period between 2005 to 2014 (European Commission, EACEA, & Eurydice 2017: 77).

In France and Germany, the expected minimum level of attainment based on the Common European Framework of Reference for Languages (hereafter CEFR; Council of Europe 2020) is B1 by the end of lower secondary, and B2 by the end of general upper secondary, whilst in Spain it is A2 and B1 respectively (European Commission, EACEA, & Eurydice 2017: 122–123). This difference in target proficiency level is reflected in the minimum annual instruction time for EFL as a compulsory subject: in the eighth year of compulsory schooling, it ranges from 111 hours in Spain to 154 in Germany and 216 in France (European Commission, EACEA, & Eurydice 2017: 107–108)[5].

Up until recently, the competence-based descriptors of the CEFR made frequent mentions of an idealised native speaker as the reference point. For instance, at B2 level, learners were expected to be able to "interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party" (Council of Europe 2001: 24) and to "sustain relationships with native speakers without unintentionally amusing or irritating them or requiring them to behave other than they would with a native speaker" (Council of Europe 2001: 76). However, the updated guidelines (Council of Europe 2020), like most current European school curricula, no longer explicitly mention native speakers as the target norm.

The German Core Curriculum stresses the need for learners of English to learn to deal with authentic texts, in particular in listening and reading comprehension, as well as in mediation activities (Kultusministerkonferenz 2012: 12, 15, 18). In Spain, too, the focus lies on the transferability of competences acquired in the classroom to genuine communicative situations:

> El enfoque orientado a la acción adoptado en el currículo se concentra en el estudiante, que es quien aprende, construye sus competencias y las utiliza, tanto para llevar a cabo las tareas de aprendizaje en el aula como las que demanda la comunicación real [The action-oriented approach adopted in the curriculum focuses on the learner, who is the one who learns, builds

[5] For Spain and Germany, these figures correspond to the weighted average instruction time as calculated on the basis of the number of students enrolled in each educational authority and type of school (European Commission, EACEA, & Eurydice 2017: 107–108).

> his or her competences and uses them, both to accomplish the learning tasks in the classroom and those afforded by real communication].[6] (Consejería de Educación, Juventud y Deporte de Madrid 2015: 133)

The French curriculum also makes clear that the priority of school English instruction should be enabling students to exploit and interact with authentic materials in all language skills (Conseil supérieur des programmes 2015).

Up until recently, there was no doubt that EFL teaching was expected to follow native-speaker norms. European foreign language curricula now generally refrain from referring to any specific native-speaker varieties. For instance, the German Education Standard for the general higher education entrance level (*Abitur*) states that:

> Sprachlicher Orientierungspunkt sind *Standardsprache(n)* sowie Register, Varietäten und Akzente, deren Färbung ein Verstehen nicht generell behindert [The linguistic point of reference is *standard language(s)*, as well as registers, varieties and accents, whose distinctiveness do not generally impede comprehension]. (Kultusministerkonferenz 2012: 14 emphasis added).

In practice, however, this typically amounts to either a 'standard' British or a US-American English norm. Using similar terminology, the curriculum of the Autonomous Community of Madrid refers to "*una variante estándar de la lengua*" [a standard variety of the language] (Consejería de Educación, Juventud y Deporte de Madrid 2015: 432). At this stage, the repeated use of the word 'standard' begs the question: What is meant by 'standard varieties of English' in such educational contexts? This question has sparked controversies since at least the 1980s, when debates concerning the variety of English to be taught as part of the National Curriculum of England and Wales raged (Tony 1999: 1). Quirk (1995: 5) provided an early, succinct summary of the main concerns associated with the term 'Standard English':

> There are few enough (not least among professional linguists) that would claim the existence of a single standard within any one of the ENL [English L1] countries: plenty that would even deny both the possibility and the desirability of such a thing. Recent emphasis has been on multiple and variable standards (insofar as the use of the word 'standard' should be ventured): different standards for different occasions for different people – and each as 'correct' as each other.

The plurality of different 'standard' registers, varieties and accents to which Quirk refers is echoed in the German Education Standard cited above – as opposed to the excerpt from the Madrilenian curriculum which uses a singular article implying that a single standard variety should be taught. Regardless of whether a single or multiple 'standards' are to be taught, in practice, what does or does not constitute a 'standard' form of any widely used language is notoriously difficult to define (for book-length discussions on Standard English, see, e.g., Crowley 2003; Milroy & Milroy 2012; Tony 1999). What most linguists, education scholars and, indeed, teaching practitioners

[6] All translations mine, unless otherwise noted.

would likely agree on, however, is that 'standard varieties' can be equated to prestige varieties (Tony 1999: 7) and prestige is usually associated with 'correctness'. This notion is confirmed in the German Education Standards, which states:

> Die Entwicklung der funktionalen kommunikativen Kompetenzen ist bezogen auf die *geläufige* und *korrekte* Verfügung über die sprachlichen Mittel in den Bereichen: Aussprache und Intonation, Orthographie, Wortschatz, Grammatik. [The development of functional communicative competence [in a foreign language] refers to the *typical/frequent* and *correct* use of linguistic features in the areas of: pronunciation and intonation, spelling, vocabulary and grammar] (Kultusministerkonferenz 2003: 9; emphases added).

Thus, in spite of not (officially) adhering to any (specific) native-speaker norm(s), the objectives set out by school educational authorities stipulate that pupils are expected to be taught *correct, typical,* and *frequent* English forms. Whilst measures of correctness necessarily involve some subjective judgements, objective measures of typicality and frequency of occurrence in English as it occurs naturally "in the wild" can be made on the basis of corpus data. At the same time, it is clear that such measures of frequency and typicality will differ depending on the situational context of language use. In sum, modern European secondary school curricula appear to advocate for the teaching of real-life, naturally occurring, idiomatic or, what has often been termed, "authentic" English.

1.3 Authenticity in EFL teaching

'Authenticity' is a particularly challenging concept in ELT and, in particular, in EFL contexts. Definitions abound – as do their interpretations (see Gilmore 2019 for an overview). One understanding, which is not infrequently encountered among English teachers, is that authentic input is input created by native speakers for native speakers (see, e.g., Little, Devitt & Singleton 2002). At the other end of the spectrum, some adopt very broad definitions such that essentially any text with a "true" communicative objective is deemed to be authentic (e.g., Swaffar 1985: 17). Since teaching and learning a language can easily be argued to constitute genuine communicative objectives, such definitions imply that all pedagogical texts are 'authentic'. In practice, this is clearly not the case: learners, teachers and researchers frequently unite to deplore the contrived, artificial-sounding texts typically found in EFL textbooks, which often feature pragmatically highly unlikely sentences of the type: *Where is Jenny, the sister of Brian?*, *Are you swimming in the sea?* [7] and *There's grass in the garden.*[8]

The real crux of the problem is that authenticity can be understood either as a characteristic of a text, the participants of the text, its communicative intent, social or cultural context, or any combination of these. Hence, authenticity need not refer

[7] From *Achievers Pre-intermediate* (see Table 5).

[8] From *Green Line 1* (see Table 5).

solely to the linguistic elements of the texts presented to learners. Indeed, some authors have (re-)defined authenticity to include the relationship of the texts offered to the learners' culture (e.g., Prodromou 1992), learners' interaction with texts and the tasks associated to them (e.g., Widdowson 1978), and the learners' personal engagement with the texts (e.g., van Lier 2013). Given this wealth of definitions, Mauranen (2004a: 201–202) suggests that it may be advantageous to distinguish between "subjective" and "objective" authenticity. Such a distinction appears reasonable at first glance. However, according to Mauranen, subjective authenticity would reflect learners' perceptions of the materials, whilst teachers' and/or researchers' evaluation would be objective. This assumption that teachers' and researchers' evaluations are inherently objective is, however, highly questionable, given that even foreign language education scholars cannot agree on an operationalisable definition.

Indeed, though space precludes a detailed discussion of the many controversies around the term and its various meanings, this very brief introduction to the issue will argue that authenticity is simply too difficult to define for anyone to be expected to make objective classifications (for more detailed discussions of authenticity in foreign language teaching and materials design, see, e.g., Widdowson 1989; Mishan 2005; Trabelsi, Tomlinson & Masuhara 2010; Tomlinson 2013a; Gilmore 2019; Nelson 2022). Part of the problem is that many of the debates on authenticity in ELT implicitly assume that a text either *is* or *is not* "authentic". Yet, authenticity need not be a dichotomous variable or, indeed, a uni-dimensional one (see also Bendix 1997: 23; Day & Bamford 1998: 58–59). That said, in the present study, the terms 'authentic', 'real-life' and 'naturally occurring' are used synonymously to refer to texts that have not been specifically produced nor modified or adapted with L2 learners in mind.
As for the pedagogical impact of authentic materials on L2 learners, some scholars have argued that the simplification and contrivance of teaching materials facilitates learning (e.g., Widdowson 1984: 218) whilst others have countered that they deprive learners of opportunities for naturalistic learning and can therefore hamper progress (e.g., Siepmann 2011: 29; Sinclair 1983; Wolff 1984). Others, still, have advocated for simplified or otherwise pedagogically modified texts that nonetheless retain the "natural qualities of authenticity" (Day & Bamford 1998: 59). The notion of 'enriched input' (also sometimes called 'flooded input') has also been proposed: it refers to pedagogical texts employed in a meaning-focused activity, in which a target structure has artificially been multiplied to raise learners' awareness of the structure in context (see, e.g., Reinders & Ellis 2009).

The most common socio-functional argument put forward by detractors of contrived pedagogical materials is that authentic materials boost learners' motivation (e.g., Ahmad & Millar 2020; Gilmore 2011; Ghanbari, Esmaili & Shamsaddini 2015; Liedke 2013; Little, Devitt & Singleton 1989; Peacock 1997; Rüschoff & Wolff 1999; Sun

2010; Varmış Kiliç & Genç İlter 2015). However, others reject this argument claiming that the opposite is true: authentic texts demotivate learners because they make too many assumptions about known lexical, grammatical and cultural knowledge (e.g., Freeman & Holden 1986; Prodromou 1996; Richards 2001: 252–254; Vielau 2005; Widdowson 2003: 107).

Given how difficult it is to define (let alone: operationalise!) authenticity, conducting valid and reliable studies to measure the impact of authentic vs. non-authentic materials on learners' motivation (or, indeed, learning outcomes) constitutes a near impossible feat. Yet, motivation remains a popular argument both for and against the use of authentic materials. The textbook publishing industry has seemingly learnt to make the best of both worlds by, on the one hand, frequently plastering claims of "authentic English" and "authentic texts" on its book covers and marketing materials and, on the other, featuring plenty of pedagogically contrived texts within its coursebooks (Gilmore 2007: 106). Perhaps the most convincing claim is that acquiring the skills to engage with authentic materials may, in itself, be motivating for learners (e.g., Little, Devitt & Singleton 1989; Skehan 2014). In other words, intrinsic motivation may arise from the process of engaging with the materials, rather than the materials themselves being the cause of the motivation (see also Gilmore 2007: 108).

At any rate, knowing that authenticity is so notoriously difficult to define certainly relativises the contentious claims made by both proponents and critics of the use of authentic materials in the EFL classroom. Hence, rather than quibble over what *does* or *does not* constitute a genuinely "authentic" or "real-life" text in the EFL classroom, or to what degree, let us focus on a more relevant and, at least theoretically, operationalisable question: What kind of language do learners need to be exposed to and engage with in order to acquire typical, idiomatic English that will equip them with the linguistic, pragmatic, discourse, and sociocultural means to thrive outside the classroom? To do so, this study turns to usage-based approaches to language and language learning (see, e.g., Barlow & Kemmer 2000; Bybee & Hopper 2001; Bybee 2007; Ellis, Römer & O'Donnell 2016; Robinson & Ellis 2008; Tomasello 2005; Tyler, Ortega & Uno 2018).

1.4 Usage-based theories to L2 learning and teaching

A usage-based understanding of L2 learning and teaching naturally draws on a range of approaches, including various branches of usage-based linguistics, foremost, cognitive linguistics (Croft & Cruse 2004; Geeraerts 2006; Goldberg 1995; 2006; Lakoff & Johnson 2003; Langacker 1987; 2008), as well as emergentism (Bybee & Hopper 2001; Ellis 1998; Ellis & Larsen-Freeman 2006; Elman et al. 1996; MacWhinney 2006), constructionism (Harel & Papert 1991; Papert 2020; Piaget 2013), and complex dynamic systems theory (de Bot, Lowie & Verspoor 2007; Fogal & Verspoor 2020;

Larsen-Freeman 1997; Verspoor 2017; Verspoor, Lowie & Van Dijk 2008). These approaches are, in turn, based on decades of research in several related disciplines, including linguistic theory, psycholinguistics, cognitive and educational psychology and, more broadly, cognitive science (for a recent overview, see Ellis 2019). At the heart of all these approaches is the central notion that: "Language and its use are mutually inextricable; they determine each other" (Ellis & Larsen-Freeman 2009: 91). Language use therefore constitutes "the foundation for language learning" (Tyler & Ortega 2018: 5).

Usage-based linguistic theories place a strong emphasis on the centrality of meaning. Meaning is, of course, also a core tenet of contemporary communicative language teaching approaches. However, in usage-based language acquisition models, the centrality of meaning implies that all aspects of language, from lexis to syntax to discourse, are acquired as form-meaning pairings. These form-meaning pairings are conceptualised as constructions of various levels of complexity and abstraction: the most concrete and specific constructions are individual words and their concrete meaning in the real world, whilst more abstract ones consist of, for example, grammatical phenomena such as the present perfect or syntactic patterns such as verb argument structures, e.g., the ditransitive construction (Ellis & Ferreira-Junior 2009a: 188; see also, e.g., De Knop & Gilquin 2016; Herbst, Schmid & Faulhaber 2014; Goldberg 2006; 1995; Hoey 2005; Lewis 2009; Siepmann 2007; Sinclair 1991 on the concept of lexicogrammar).

Language is acquired as a result of exposure to these form-meaning mappings, or constructions, in "iterative usage events" (Tyler & Ortega 2018: 7; see also Barlow & Kemmer 2000) – in other words, when language users are exposed to, engage with, and produce surface-level linguistic patterns that convey specific meaning in genuine communicative situations. It is through these usage experiences that linguistic knowledge becomes entrenched in the learner's mind (see, e.g., Blumenthal-Dramé 2012). As such, and contrary to generative, rule-based theories of language acquisition (as postulated by, e.g., Chomsky 1995; 2002), humans are not endowed with any universal or innate abstract grammar rules. Consequently, language acquisition processes are not concerned with the setting of any parameters for pre-supposed innate grammar rules, but rather language users naturally construct rules as they gradually make generalisations on the basis of the linguistic patterns they encounter over time. This, in turn, means that language structure – in the form of various kinds of constructions, e.g., words, collocations, grammar, and discourse – cannot be successfully acquired if dissociated from meaning. Hence, following a usage-based approach to language acquisition, pedagogically contrived textbook texts and contextless sentences exemplifying grammar rules are not thought to be the most successful means to learn a language.

Language learning is a process and, in this paradigm, the ability to generalise over individual language usage events and to induce meaningful categories are understood as examples of domain-general cognitive processes, i.e., abilities that are not specific or restricted to language learning (e.g., Cohen & Lefebvre 2017; Murphy 2003), but which are rather at the heart of all aspects of human learning. For language acquisition, these general cognitive mechanisms (such as memorisation, pattern finding, abstraction, induction, categorisation and schematisation) have been shown to be driven by various aspects of the input language users are exposed to; in particular, they are known to be exquisitely sensitive to frequency effects (e.g., Ellis 2002). Indeed, the degree of entrenchment of any one construction in the learner's mind is thought to be proportional to the frequency of usage (e.g., Bybee & Hopper 2001; Bybee 2007; Ellis 2002; Tomasello 2005). This means that, in a usage-inspired L2 instruction paradigm, both the quantity and the quality of language input are crucial.

1.5 Input and frequency

Input is, in fact, central to all theories of L2 learning and teaching and: "no model of second language acquisition does not avail itself of input in trying to explain how learners create second language grammars" (Gass 1997: 1; for more on the "input hypothesis" in SLA, see Krashen 1982; 1985). In usage-based accounts, input is understood as a wealth of information that captures both the frequencies at which various linguistic patterns occur in natural usage and the contexts in which they are most likely to (co-)occur (Bybee & Hopper 2001; Bybee 2007). Numerous psycholinguistic, corpus- and computational-linguistic studies have now demonstrated that "the acquisition of constructions is input-driven and dependent on learners' experience with form-meaning mappings in context" (Ellis & Larsen-Freeman 2009: 92). For instance, experiments (e.g., Elio & Anderson 1981; 1984; Posner & Keele 1968; 1970) have shown that the learning of categories from prototypical exemplars is optimised when learners are exposed to language input in which the distribution of specific exemplifications of a construction are heavily skewed towards one prototypical exemplar (e.g., the verb *give* in ditransitive constructions). As they are exposed to more input, learners continually redefine the bounds of each construction category (Ellis & Larsen-Freeman 2009: 95). Corpus studies (e.g., Ellis, Römer & O'Donnell 2016; Goldberg, Casenhiser & Sethuraman 2004) have demonstrated that the natural distributions of constructions follow Zipf's law (Zipf 1935; 1949), whereby one exemplar is by far the most frequent and the frequency of the second most frequent exemplar can be expected to be approximately half the frequency of the first exemplar and so on (for a succinct explanation with clear illustrations, see Brezina 2018: 44–46). These highly frequent, semantically prototypical exemplars are thought to serve a 'pathbreaking' function facilitating the generalisation of a construction to more abstract instantiations (Goldberg 2006; Ninio 1999).

These effects have been demonstrated in both L1 and L2 acquisition contexts (e.g., Goldberg, Casenhiser & Sethuraman 2004; Lavi-Rotbain & Arnon 2022; Römer & Berger 2019). For instance, Ellis and Ferreira-Junior (2009a; 2009b) analysed ESL adult learners' use of three types of verb-argument constructions over a period of 32 months and found that participants' first use of each verb-argument construction was indeed with a 'pathbreaker' verb, which corresponds to the most frequent verb type of each construction in the learners' language input. Moreover, these pathbreaker verbs appear to "seed" the construction so that, over time, learners begin to use more (semantically similar) verbs as part of the process of mental abstraction of the construction. The results of these studies (see also Wulff et al. 2009) suggest that the naturally skewed distributions of construction type/token frequencies in natural language input optimises not only L1 but also L2 learning by providing one very high-frequency exemplar that is also prototypical in meaning and widely applicable in a broad range of contexts (cf. Hu & Maechtle 2021).

Thus, usage-based linguistic theories and the results of empirical studies converge to show that language learning is largely driven by frequency. That said, it is not by any means the only factor that contributes to successful language learning. Just like in all other non-linguistic learning processes, socio-emotional factors are also considered in usage-based approaches to language learning. These include surprise value, learner attention, transfer, overshadowing and blocking (see, e.g., Ellis 2002; 2006; 2008). Other important factors include the salience of a construction to comprehend or produce a particular utterance, as well as the prototypicality, generality or redundancy of a construction. Suffice to say, however, that, in naturalistic input, these aspects very often, though by no means always, correlate strongly with frequency of use. For instance, the most frequent forms are unlikely to be redundant and, as explained above, it is typically the most frequent exemplar of a construction that gives that construction its prototypical meaning (e.g., *give* in ditransitive constructions).

1.6 Input in lower secondary school EFL contexts

As we have seen, input is central to almost all SLA models and, in particular, to usage-based approaches, yet empirical research on the impact of input on L2 learning development remains relatively sparse (Gurzynski-Weiss et al. 2018: 292). Part of the issue is that, whilst the role of input in early L1 acquisition can be relatively easily examined by analysing the language of a young learner's main caretaker language (see, e.g., Altosaar et al. 2010; Behrens 2006; Clark & Casillas 2016; Kuhl & Meltzoff 1996), capturing the language L2 learners are exposed to is rather more complex:

> Rather than a single or limited set of caretakers, second language (L2) learners are exposed to numerous native and nonnative speakers, making it nearly impossible to accurately characterize all sources of input (Gurzynski-Weiss et al. 2018: 291).

The few studies that have attempted to do so have, however, concluded that input is equally important in L2 as in L1 acquisition (see, e.g., Moyer 2008 on phonological attainment). Whilst capturing the total language input of (adult) learners in ESL contexts is particularly tricky, in lower secondary school EFL contexts, L2 input is rather more restricted and thus easier to capture. At lower secondary school level in France, Germany and Spain extracurricular exposure to and interaction with English remains, on average, fairly limited (Berns, De Bot & Hasebrink 2007). Note, however, this is not the case everywhere (see, e.g., Henry 2014 for the case of Sweden) and that the situation is rapidly evolving. Indeed, recent surveys have shown that the proportion of teenagers in Germany who consume English-medium media is on the rise: 43% of 12- to 19-year-olds in Germany report watching YouTube videos in a language other than German – mostly English – at least once a week (Feierabend et al. 2020: 48). The percentages reported at lower secondary school level are only marginally lower: 32% for 12–13-year-olds and 39% for 14–15-year-olds (Feierabend et al. 2020: 48). We can expect similar trends in France and Spain: as teenage EFL learners' English proficiency grows, so does their consumption of media in English and engagement in English-medium (online) communication – making it increasingly difficult to discern how much of their L2 input is classroom-based.

At the time of writing, however, formal classroom-based English input remains the dominant source of English input for most pupils at lower secondary school level in the three countries of interest (more on this in 1.7). This input consists foremost of the content of the textbook (that is: the student's coursebook, associated audio and video materials and potentially also a workbook and/or vocabulary book), teacher talk, peers' production, and, if used, any additional teaching materials. The present study was thus motivated by the combination of the centrality of input in L2 acquisition processes and the fact that a substantial proportion of that input in lower secondary school EFL contexts comes from textbooks. Indeed, although other sources of classroom-based L2 input have just been listed, large proportions of these sources are in fact directly or indirectly influenced and/or mediated by textbook content. For instance, much of teacher talk at secondary school level revolves around the textbook, its explanations, instructions, and tasks, and much of learner writing, learner-teacher and learner-learner spoken interactions are produced on the basis of these same textbook tasks, prompts and models (see, e.g., Huang 2019: 87; Thornbury 2002).

1.7 Textbooks in the EFL classroom

Although statistics are hard to come by (Schaer 2007: 255), there is a broad consensus on the fact that textbooks largely dominate formal L2 input in European lower secondary school EFL contexts. In fact, textbooks are almost universally "considered to be the backbone of second and foreign language teaching" (Tateyama 2019: 404; see also, e.g., Diepenbroek & Derwing 2014; Oelkers 2008). Richards (2015: 594) goes as far claiming that they largely determine teachers' teaching practice. Across all EFL

instructional contexts, it has been deplored time and again, by language education scholars and teacher trainers alike, that teachers are heavily dependent on textbooks and that there is *de facto* no distinction between textbook and syllabus (e.g., Schaer 2007: 256; Sinclair & Renouf 1988: 145). According to Vellenga (2004: n.p.; see also, e.g., Hyland 1994; Kim & Hall 2002), they constitute both "the centre of the curriculum and syllabus in most [EFL] classrooms". Thornbury (2012 in a comment response to Chong 2012: n.p.; see also Bragger & Rice 2000: 107) goes further and claims that:

> the reason that coursebooks are so often in the line of fire is that they do to a large extent dominate and determine so many aspects of a teacher's day-to-day professional life. They (more often [than] not) instantiate the curriculum, provide the texts, and – to a large extent – guide the methodology.

This view is echoed in the following observations:

> Together with teaching methodologies, [textbook] materials represent the interface between teaching and learning, the point at which needs, objectives and syllabuses are made tangible for both teachers and students. They provide most of the input and language exposure that learners receive in the classroom […]. (Hyland 2013: 391)

As Usó-Juan & Martínez-Flor (2010: 424) stress, textbooks have always tended to be "[t]he main source of input presented in classroom settings" (see also Tono 2004: 45); hence, this is not a new phenomenon but, in spite of much criticism of textbooks, it is also not one that appears to be changing in any significant way.

Virtually all European lower secondary EFL classrooms are equipped with textbooks. Thus, the overwhelming dominance of the textbook as a source of L2 input that has been observed in general, global EFL contexts is likely to be also true of lower secondary school EFL education in mainland Europe. Regarding the secondary school German context specifically, Kurtz (2019: 116) speaks of the "großenteils lehrwerkorientierte Alltagspraxis des Englischunterrichts [largely textbook-oriented everyday practice of teaching English]", especially in the "Sekundarstufe I" (Kurtz 2019: 122; see also Hermes 2009: 9), i.e., in the first five years of secondary school. Similarly, Volkmann (2010: 235) reports that:

> das traditionelle Leitmedium des Unterrichts, das Lehrwerk, insbesondere das Lehrbuch (Schülerbuch), bleibt in der Phase des Spracherwerbs (also vor allem in der Sekundarstufe I) das oftmals absolut dominante Medium der Instruktion) [as the medium which has traditionally guided and organised teaching, the textbook, especially the coursebook (i.e., the pupil's book), often remains the absolutely dominant medium of instruction in the language acquisition phrase (i.e., especially at lower secondary school level)].

In some *Bundesländer*, this reliance on textbooks is, in fact, more or less directly prescribed in the curriculum. The English curriculum for *Gymnasium* in Hessen, for instance, proclaims that, at lower secondary level (*Sekundarstufe I*) the textbook is *the* "Leitmedium" (Hessisches Kultusministerium 2010: 4).

Referring to the German context more generally, Siepmann (2007: 59) points to a noticeable overlap between textbooks, syllabus and vocabulary teaching methodology when he notes that:

> [i]n der Sekundarstufe I verlassen sich die Lehrer auf das Wörterverzeichnis und die Grammatik des Lehrbuchs; die ausgeprägte Einzelwortorientierung dieser Lernhilfen wird im Unterricht übernommen [at lower secondary school level, teachers rely on the textbook's vocabulary and grammar sections; the strong emphasis on individual words that these materials promote is mirrored in vocabulary teaching and explanation].

Over in Spain, Alejo González et al. (2010: 61) make very similar observations:

> Spanish secondary school students meet English first and foremost in the language classroom and the coursebook that they use is likely to be their primary source of English language input.

The situation in France is particularly interesting as teachers' perceived over-reliance on textbooks has led to calls to abandon textbooks altogether or, at the very least, to adapt and/or supplement textbook materials with 'authentic' texts (see 1.3). In particular, teachers in their post-studies qualification stage are often told by teacher trainers and assessors to avoid relying on a textbook in their observed (and assessed) classes. This backlash has led many French EFL teachers to resorting to a mix-and-match approach – combining texts and activities from several textbooks and additional resources, rather than relying on a single textbook series (personal communication with practising teachers, see also Various users of neoprofs.org 2016; Séré & Bassy 2010: 10–11). Nonetheless, even when surveyed by the Ministry of National Education, French EFL teachers report that, at *collège* level [lower secondary school, see Table 3], textbooks are "*indispensable*" (Leroy 2012: 62).

The palpable tension concerning the use of textbooks in French EFL classrooms is, in fact, symptomatic of a far more universal love-hate relationship with textbooks in ELT. Indeed, in spite of the undeniable popularity of EFL textbooks, as demonstrated by their widespread use in foreign language classrooms across the world and the great range of publications on offer, detractors have regularly deplored the "superficial and reductionist" content of textbooks, that impose "uniformity of syllabus and approach" and remove "initiative and power from teachers" (Tomlinson 2001: 67). Following this line of thought, Prabhu (1989) argues that textbooks rob teachers of the freedom to freely order, use and localise materials.

Another important factor to consider is that most EFL textbook publishing houses are commercial, for-profit businesses. Some have therefore claimed that textbook publishers do not always have learners' best interests at heart since, at the end of the day, learners are rarely involved in textbook selection processes (in fact, it is not rare for teachers to be entirely excluded from textbook selection processes, too, see, e.g., Friederici 2019; Stein et al. 2001: 5–6; Stranks 2013: 338). It has been argued that privately outsourcing such a crucial aspect of EFL education has the potential to stall

the implementation of recent research findings in applied linguistics, second language acquisition and other relevant disciplines. Some claim that "the economic imperative" incites publishers to "clone previously best-selling coursebooks rather than risk investment in more principled innovations" (Tomlinson 2013b: 541; see also the conclusions of Burton 2019 whose analysis of the "canon of pedagogical grammar" in ELT textbooks combines textbook content analysis and interviews with textbook authors, editors and publishers). The situation can be summarised as follows:

> Publishers obviously aim to produce excellent books which will satisfy the wants and needs of their users but their need to maximize profits makes them cautious and conservative and any compromise with authors tends still to be biased towards perceived market needs rather than towards the actual needs and wants of learners (Tomlinson 2013b: 3).

This is not to say that attempts at innovation have not been made – on the contrary. For example, the 'lexical approach' (Lewis 1993; 1997; 2009), that challenges the all-empowering centrality of grammar in the L2 syllabus, inspired the design of a number of commercially published textbooks (e.g., Dellar & Hocking *Innovations*, 2000; Dellar & Walkley, *Outcomes*, 2011). These emphasised the importance of conceptualising language as 'grammaticalised lexis' as opposed to the customary 'lexicalised grammar' approach (Lewis 1993: 34) using corpus-informed texts and activities. Presumably these were not great commercial successes, however, because a brief tour of the tables of contents of today's most popular EFL textbooks clearly shows that these continue to treat grammar and vocabulary as two distinct areas of language teaching and learning (see also Tan 2003). Again, it is easy to see how school textbook publishers would be placing themselves at a competitive disadvantage if they were the first to remove what has come to be an expected feature of foreign language textbooks and has, so far, proved to be an attractive selling point.

The constraints associated with the commercial production of ELT materials have been extensively discussed in, among others, Bell & Gower (2011), Richards (2015) and Gray (2010). Although these publications tend to focus on the global ESL/EFL textbook market, most often targeted at adult learners, continental European publishers producing school textbooks for their respective domestic markets likely face many of the same constraints.

An additional constraint, and one that may be more specific to the European textbook market, is that European foreign language curricula and syllabi are now largely aligned with the CEFR. Indeed, the CEFR has established itself as an unavoidable pedagogical framework for language learning and teaching in European schools and, as such, has had a major influence on textbook and task design (Hallet & Legutke 2013: 8) in spite of much criticism of the framework (for the German context, see, e.g., Bausch 2005; Vogt 2011).

Given the widespread criticism of textbooks for a multitude of reasons, we may ask: why *are* textbooks nonetheless so ubiquitous in ELT and even considered "indispensable" (Leroy 2012: 62) by many secondary school EFL teachers? As any practising teacher can attest, textbooks are, first and foremost, a much-needed timesaver when teachers' timetables are packed and classes full (see, e.g., Nordlund 2016: 48). Recognising a genuine need to reduce teachers' preparation load, textbook publishers have responded by adapting their business models and are now marketing all-encompassing "multidimensional packages" (Dat 2013: 409) which go well beyond what was traditionally understood as a textbook. For each textbook within a textbook series, these packages now frequently include, in addition to the pupil's coursebook, an activity workbook and a teachers' manual with often very detailed lesson plans, step-by-step instructions, extra photocopiable worksheets, answer sheets, as well as optional related games, quizzes, and assessments, vocabulary apps, audio recordings, videos, graded readers, etc. (Dat 2013: 410–411). Given the wealth – and, it is worth highlighting, the often high quality – of these materials, it is easy to see how, in particular inexperienced, teachers can quickly come to rely on them so much.

In general, the textbook (package) is perceived as a trustworthy authority (e.g., Abello-Contesse & López-Jiménez 2010; Brown 2014; Chien & Young 2007; Ghosn 2013). There is often a sense that, if it is followed to the letter, teachers can be reassured that their lessons will cover all aspects of the curriculum and syllabus (Nordlund 2016: 48). Textbooks therefore contribute to standardising learning outcomes (Anton 2017: 13). In many cases, textbooks also act as a mark of credibility vis-à-vis ever-more demanding parents. Furthermore, textbooks are frequently seen – by learners, teachers and parents alike – as an ideal way to present contents in a well-structured and systematically organised order, following tried-and-tested progressions (Burton 2019; 2020; Möller 2016). This leads to:

> a circle (whether vicious or virtuous), whereby publishers provide their customers with the kind of teaching materials that they are asking for, and their customers continue to ask for the same kinds of teaching materials as they feel that what they have seen before represents the norms they should be following (Burton 2019: 220–221).

The fact that these norms and progressions may be the product of decades of innovation stagnation rather than the conclusion of any empirical studies on learners' development of linguistic competence in instructional EFL settings is usually overlooked. On the contrary, textbooks are often perceived as "Innovationsträger [drivers of innovation]" that bring pedagogical research findings and new teaching methods to the foreign language classroom (Anton 2017: 14).

Given the rapid growth in technology-based ELT and, more generally, computer-assisted language learning (CALL; see, e.g., Chapelle 2010), it may seem rather inconceivable that, in the late 2010s and early 2020s, secondary school learners' main source of formal English input still comes in the form of book publications (Bezemer

& Kress 2016: 477). Vague claims and slogans such as "*Die Perspektive des Schulbuchs ist digital* [The future of the textbook is digital]" (Landesregierung Nordrhein-Westfalen 2016: 8, 25) found in North-Rhine Westphalia's *Leitbild 2020 für Bildung in Zeiten der Digitalisierung* [Mission statement for 2020 for education in the day and age of digitalisation] are, indeed, frequently heard. Yet, to date, the vast majority of so-called "digital textbooks" and their accompanying e-materials are essentially replicas of the same textbooks, graded readers, grammar books and flashcards that publishers still successfully sell in paper form. Hence, although all major school textbook publishers now promote various digital textbook packages (Kurtz 2019: 119), for now, these digital textbooks offer little more than digitised versions of their paper counterparts. They represent little to no change in terms of content or teaching methodologies (see, e.g., Gehring 2013; Richards 2015: 594; Stranks 2013: 348–349; Schildhauer, Schulte & Zehne 2020: 30–31). The obvious lack of suitable digital materials (as well as, crucially, teacher training in using existing digital resources and, in many cases, the necessary equipment and infrastructure) made headlines during the (partial) school closures triggered by the COVID-19 pandemic (see, e.g., Blume 2020; Fominykh et al. 2021; Kerres 2020; Starkey et al. 2021; van de Werfhorst, Kessenich & Geven 2020). Whilst the urgency of the epidemiological situation is likely to have accelerated both the development and the acceptance of new digital teaching materials compatible with online, on-site and hybrid instructional settings, genuine advances in commercial materials development can nevertheless be expected to remain slow. Indeed, academic research has been churning out innovative, evidence-based ideas for new digital L2 teaching materials for the better part of a decade (see, e.g., Biebighäuser, Zibelius & Schmidt 2012; Meurers et al. 2010; 2019); however, few of these ideas have been translated into any of the best-selling secondary school EFL textbook series examined as part of the present study. It would thus appear that textbook publishers face the same constraints as ever. In the textbook industry, innovation remains a commercial risk.

As this section has shown, Vellenga's (2004: n.p.) statement that "textbooks remain the most important tools and resources in the EFL classroom" still rings true today – in spite of their many shortcomings (see also Möller 2016). In addition, it has concluded that textbooks continue to play a particularly important role at lower secondary school level – accounting for a substantial, if not the largest, proportion of L2 input EFL learners are exposed to. Gaining a comprehensive understanding of the language that modern secondary school EFL textbooks present to learners is therefore of high pedagogical value. This is precisely what the present thesis sets out to achieve.

The approach adopted to do so is corpus-based. In other words, the totality of texts from a representative sample of EFL textbooks used at lower secondary school level in France, Germany and Spain (see Table 5) is analysed as a "learners' L2 [input] corpus" (Gabrielatos 1994: 13; see also Meunier & Gouverneur 2007: 122). Thus,

rather than following a page-by-page textbook analysis approach, the language of this corpus of "Textbook English" (see 3.3.1 for details of its composition) is examined as a variety of English, much like Academic English, Australian English or Aviation English.

1.8 Corpus linguistics and foreign language education

The present corpus-based textbook analysis study follows in the footsteps of a now decade-long tradition of "pedagogy-driven corpus-based research" (Gabrielatos 2006: 1). Corpus-based methodologies rely on the exploration of language corpora, principled computerised collections of real-life, authentic texts, to investigate patterns of language use. Corpus linguistics is characterised by its empirical basis, analysing (usually large) collections of texts using automatic and interactive data retrieval techniques, and by its application of mixed quantitative and qualitative analytical methods. Drawing on corpus data without (too many) assumptions allows linguists to observe language features, e.g., lexico-grammatical patterns and other phenomena, which have not necessarily been previously explored or described (Hunston 2002: 1). For example, analyses of corpora of spoken British English revealed highly frequent lexico-grammatical features in spoken English which had previously not been considered in traditional grammars (see, e.g., Carter & McCarthy 1995; McCarthy & Carter 1995; Carter, Hughes & McCarthy 1998; Hughes 2010; McCarthy 1998).

As a discipline, corpus linguistics has, from the outset, positioned itself as a decisively applied subdiscipline of linguistics. Pedagogical applications have been at the heart of many strands of corpus-linguistic research and corpus methods[9] are now widely used in numerous areas of applied linguistics relevant to second language acquisition and foreign language education. In particular, corpus-linguistic methods have now become the norm in (learner) lexicography (see, e.g., Granger 2018; Rundell 2008; Runte 2015) and, since the 1990s, have had a major impact on the development of reference and learner grammars of English (e.g., Biber et al. 1999; Conrad, Biber & Leech 2011 for English; see also Siepmann 2018a; 2019; Siepmann & Bürgel 2022 for a corpus-based learner grammar of French). For instance, the second edition of the *Longman Grammar of Spoken and Written English* (Biber et al. 1999) relied exclusively on empirical data drawn from corpus analyses (Conrad 2000: 548–549). Moreover, corpora of learner language, both written and spoken, have been used in contrastive studies comparing learner language to native and/or non-native expert language use to investigate the influence of learners' L1s on their L2 productions (e.g., Bruyn & Paquot 2021; Granger, Hung & Petch-Tyson 2002; Tracy-Ventura & Paquot 2020), as well as in a host of natural language processing (NLP) applications, e.g., to automatically score and mark learner texts, perform proficiency level classification,

[9] For more on the debate of corpus linguistics as a discipline vs. a methodological framework, see Stefanowitsch (2020: 21–60) and Taylor (2008).

error detection and/or correction (e.g., Ballier, Díaz Negrillo & Thompson 2013; Leacock et al. 2010; Meurers 2015; Reder, Harris & Setzler 2003).

According to Granger (2004: 136), the main fields of pedagogical application of corpus data are classroom methodology and materials and syllabus design. However, Granger (2004: 136) adds that "with the exception of ELT dictionaries, the number of concrete corpus-informed achievements is not proportional to the number of publications advocating the use of corpora to inform pedagogical practice". Recent studies appear to confirm that this statement is, unfortunately, very much still valid today (see, e.g., Callies 2019; Chambers 2019; Jablonkai & Csomay 2022).

Concerning the impact of corpus data on pedagogical methods, most research has so far focused on data-driven learning. However, in spite of the wealth of publications on data-driven learning going back to the work of Tim Johns from the 1980s onwards (e.g., Johns 1986; 1993; 2002; 2014) and a myriad of studies pointing to its effectiveness in a wide range of teaching contexts (summarised in two recent meta-analyses Boulton & Cobb 2017; Lee, Warschauer & Lee 2019), the direct use of corpora in the foreign language classroom has yet to become more than an exception to the norm (Barbieri & Eckhardt 2007: 320; see also Callies 2019; Leńko-Szymańska & Boulton 2015; Leńko-Szymańska 2017; Mukherjee 2004).

As for the application of corpus data in materials and syllabus design, corpus linguists have long sung the merits of incorporating corpus-based findings in L2 materials in a way that will inevitably require some modifications to traditional foreign language syllabi (e.g., Biber & Reppen 2002a; Conrad 2000; Frazier 2003; Harwood 2005; Holmes 1988; Granger 2004; McCarthy & Carter 1995; Nelson 2022; Timmis 2013); yet, in spite of the growing availability of freely accessible corpora and corpus research findings, very few EFL textbooks are advertised as corpus-informed, let alone corpus-based. In the rare cases where corpora do inform EFL textbook design, it tends to be in the context of English for Special Purposes (ESP) and English for Academic Purposes (EAP) textbooks (Meunier & Gouverneur 2009: 180–181). General EFL textbooks, by contrast, appear to remain largely unaffected by such moves (for a notable exception see the *Touchstone* series by McCarthy et al. 2006).

When Prowse asked ELT materials designers how they approached textbook writing back in 1988, the authors stressed the creative nature of the writing process. Prowse (Prowse 1998: 137) concluded that most textbook authors:

> appear to rely heavily on their own intuitions viewing textbook writing in the same way as writing fiction, while at the same time emphasizing the constraints of the syllabus. The unstated assumption is that the syllabus precedes the creation.

A few decades later, Burton (2012) conducted a case study survey of fifteen EFL coursebook authors, which revealed that authors still largely relied on their intuition.

Accessibility issues, lack of relevant skills and knowledge, and time constraints were all cited as reasons for their lack of use of corpora designing ELT materials. Given the wealth of English-language corpora and accessible, user-friendly tools that became available over the past few decades, this lack of innovation is regrettable. Indeed, as corpus-based English grammars such as the *Longman Grammar of Spoken and Written English* (Biber et al. 1999; see also Biber et al. 2021 for a more up-to-date corpus-based English learner grammar) have since shown:

> Unfortunately, decisions about the sequencing of material, typical contexts, and natural discourse are not served as well by intuition and anecdotal evidence as judgments of accuracy are (Biber & Conrad 2001: 1).

Analysing in-depth interviews with four ELT editors employed by Cambridge University Press (CUP), Curry et al. (in press) yielded more recent insights into what textbook editors currently perceive as the advantages and limitations of corpus linguistics for ELT materials development. It transpired that the perceived limitations are largely traceable to limited knowledge about existing corpora (including what kind of corpus metadata are available and how they can be exploited) and corpus tools.

The conclusions of this most recent survey are particularly sobering considering that CUP likely represents a notable exception in the ELT publishing world; indeed, it has a long tradition of collecting and processing data for the (co-)development of language corpora. Most notably, it has been instrumental in the development of the Cambridge Learner Corpus, which is used by CUP authors to target common learner errors in ELT publications, including textbooks. In this respect, it also constitutes an exception to what Granger (2015: 494) describes as learner corpora's "more nominal than real" impact on textbooks.

If corpora and the insights of corpus-linguistics studies have yet to be taken on board by EFL textbook authors, editors and publishers, it is nonetheless possible to examine and evaluate the language of textbooks using corpus-linguistic methods (see also Nelson 2022). This is what the present thesis sets out to do. As will be shown in the following literature review chapter, it is not, by any means, the first study to attempt to do so. The organisation of this state-of-the-art chapter, as well as of the remaining chapters of this thesis is spelt out in the following section.

1.9 Outline of the thesis

Having explained the background to and motivation behind the present study, the following literature review chapter (Chapter 2) provides an overview of state-of-the-art research on the language of school EFL textbooks. It is divided into two parts. Part 1 is a methodological review in which the various methods employed so far to analyse, describe and evaluate Textbook English are explained and illustrated with

selected studies. Part 2 summarises the results of existing studies on various aspects of Textbook English, including lexical, grammatical and pragmatical aspects. Based on the methodological limitations and the gaps identified in the existing literature, Chapter 3 elaborates the specific research questions addressed in the present study. These research questions informed the decision-making processes involved in the compilation of the Textbook English Corpus (TEC) and the selection/compilation of three target learner language reference corpora. These processes and their motivations are explained in the remaining sections of Chapter 3.

The research questions are then tackled at two levels. First, Chapters 4 and 5 examine Textbook English at the micro-level – focusing on two individual lexico-grammatical features in two case-study chapters. Chapter 4 analyses a grammatical feature explicitly taught in the grammar sections of the examined textbooks – the progressive, whilst the following chapter turns to the more implicit use of specific lexico-grammatical patterns involving the high-frequency verb MAKE. Chapters 6 and 7, by contrast, explore Textbook English at the macro-level, using multi-variable dimension-reduction statistical methods to gain a more comprehensive understanding of the linguistic nature of Textbook English on multiple dimensions of variation.

Chapters 4 to 7 each rely on different corpus-linguistic methods and therefore each of these four results chapters begins with its own methods section. Parts of Chapter 4 are essentially an extended conceptual replication of Römer's (2005: Ch. 5–6) extensive analysis of the representations of the progressive in the dialogues of EFL textbooks used in German secondary schools with some methodological refinements. Section 4.2.3 explains the underlying principles of the family of collostructional analysis methods (Stefanowitsch & Gries 2003; Gries & Stefanowitsch 2004a; Gries 2019a) as these are applied in both Chapters 4 and 5. Chapter 6 describes Textbook English using Biber's (1988) multi-dimensional model of General Spoken and Written English. Chapter 7 features the most extensive methods section as it proposes a range of modifications to the well-established multi-feature/dimensional analysis framework (Biber 1988; 1995; Berber Sardinha & Biber 2014; Berber Sardinha & Veirano Pinto 2019) based on limitations of the method observed in Chapter 6.

Using this revised multi-feature/dimensional analysis (MDA) framework, Chapter 7 presents the results of two multi-dimensional models of Textbook English which, across several major dimensions of linguistic variation, explain, on the one hand, the various sources of linguistic variation within EFL textbooks and, on the other, the way in which Textbook English is both, in some respects, similar to and, in others, different from the kind of English that EFL learners are likely to encounter outside the classroom. Thus, the models contribute to a much better understanding of the linguistic specificities of Textbook English which, in turn, has implications for teachers, textbook authors, editors, publishers and policy-makers. For instance, with

a greater awareness of how far-removed textbook dialogues are from natural conversation, teachers may want to supplement textbook materials with other teaching materials that better represent real-life conversation.

Chapter 8 summarises and brings together the results of the two case studies and the insights from the new multi-feature and multi-dimensional models of Textbook English presented in Chapter 7. Both methodological and pedagogical implications are discussed. Lexico-grammatical aspects of Textbook English that substantially diverge from the target reference corpora are highlighted. Suggestions are made as to how teachers, textbook authors and editors may improve unnatural-sounding pedagogical texts. Suggestions are also made for triangulating the results with learner corpus data to investigate the impact of Textbook English on EFL learners' productive competences. Some of the corpus-linguistic methods applied in the present study may also be of interest to textbook authors, editors, publishers, and representatives of education authorities interested in evaluating and/or comparing the suitability and naturalness of pedagogical texts. To this end, and in the interest of Open Science (see 3.2.2), the data and code for all the analyses carried out as part of this project have been published in a public repository on GitHub (https://github.com/elenlefoll/TextbookEnglish). They are also accessible from the Online Appendix hosted on https://elenlefoll.github.io/TextbookEnglish.

2 Literature review

Tradition, even if it is most venerable,
cannot serve as a substitute for research.
❖ Dieter Mindt (1997a: 41)

Having established that both the frequency and quality of input is fundamental to L2 acquisition and that, at least in the context of European secondary schools, textbooks account for a large proportion of learners' language input, the following question arises: Are secondary school textbooks providing the kind of language input that will promote 'authentic' language acquisition, or, to quote the Standing Conference of the German Ministers of Education (Kultusministerkonferenz 2003: 9) "the correct use of typical and frequent linguistic elements"? In an attempt to shed light on this question, the present chapter will present previous research on the language of EFL textbooks. Only where methodological innovations or specific fields have been left out in English for General Purposes (EGP) or secondary school EFL textbook studies, will the occasional reference to textbooks of other foreign languages, or textbooks for other levels and learning contexts be made.

Textbooks have long been a cherished object of study in a wide range of disciplines applying an equally diverse array of methods. As "social-cultural-political artefact[s]" (Singapore Wala 2013: 120), foreign language textbooks may also be considered as "sources not only of grammar, lexis, and activities for language practice, but, like Levi's jeans and Coca Cola, commodities which are imbued with cultural promise" (Gray 2000: 274). As such, it is quite natural that the broad spectrum of EFL textbook studies should include fields of research as disparate as the pragmatics of politeness in German EFL textbooks (Limberg 2016), semiotic approaches to the representation of culture in Hungarian EFL textbooks (Weninger & Kiss 2013), and the evaluation of interactional metadiscourse in Iranian EFL textbooks (Alemi & Isavi 2012), using an equally broad range of different methods. Though research on EFL textbooks extends well beyond "the linguistic nature of their content" (Littlejohn 2011: 182), the present study focuses exclusively on the language of textbooks – as opposed to the pedagogical reasoning behind the textbooks' tasks and activities and their effectiveness (see, e.g., Harwood 2005; Jacobs & Ball 1996; Ranalli 2003), its layout or the nature of content topics chosen by the textbook authors (see, e.g., recent special issue of *Language, Culture and Curriculum*; Canale 2021; also Siegel 2014), or its adherence and fulfilment of specific educational standards (e.g., Cools & Sercu 2006 on the extent to which the tasks and topics of two German as a Foreign Language textbooks are aligned with the CEFR) – hence the studies reviewed in the present chapter all focus on the linguistic content of EFL textbooks, hereafter referred to as 'Textbook English'.

The present literature review does not claim to cover the full breadth of past and current research on the language of EFL textbooks. In fact, to the author's best knowledge, no systematic review of Textbook English studies has been attempted so far. This is likely due to the incredibly diverse range of methods and linguistic foci that characterise this field of study, as well as the many different types of English textbooks that cater for different instructional settings, proficiency levels and regional markets and are therefore not readily comparable. The most comprehensive overview of English textbook research to date can be found in Meunier & Gouverneur (2009: 183–184). A total of 27 studies, spanning from 1990 to 2009, are summarised in a tabular format. The overview not only covers the linguistic content of textbooks, e.g., its "authenticity", grammar and vocabulary, but includes a few studies on non-linguistic aspects of textbook research, e.g., task design.

Following a similar approach, all the relevant studies surveyed as part of this literature review are summarised in a table that can found in the Online Appendix 2.1. It presents the results of a non-exhaustive survey of Textbook English studies published over the past four decades, summarising some of the key information on each study, including its main language focus, methodological approach, information on the textbooks investigated, and, if applicable, on any reference corpora used. Empty cells represent fields that are either not applicable to this particular study or for which no information could be found. Intended as a dynamic resource, this interactive, searchable, and filterable table currently lists over 80 studies on the language content of English L2 textbooks, thereby demonstrating the breadth of Textbook English studies published to date.

In light of the sheer number of publications on the subject, the present chapter can only aim to provide key insights from a selection of studies. To this end, this chapter is subdivided into two main parts. Part one (2.1) focuses on the methodologies applied from the 1980s to the present day with the aim of investigating the authenticity and/or pedagogical relevance of various features of Textbook English. Summaries of individual studies serve to exemplify the methodological approaches described. In this first part, the results of studies are only presented to illustrate the advantages and limitations of each method. Part two (2.2) then reports on key results from a range of relevant Textbook English studies, including the ones outlined in the methodological part of the chapter (2.1). Since some aspects of Textbook English have been at the heart of more than one study, this second half is organised in sections that roughly correspond to the different types of linguistic features examined in these studies (tense, aspect, lexis, etc.), rather than by chronological order. The chapter concludes with a list of implications for the present study that concern both the choice of data and methods and the language focus of the ensuing analyses (2.3).

2.1 Part One: Methodological review

2.1.1 Intra-textbook approaches

The studies outlined in the following sections focus exclusively on describing Textbook English without relying on any form of comparison with other sources of English. Such 'intra-textbook' approaches are illustrated in the following sections with, first, check-list approaches (2.1.1.1), second, largely qualitative page-by-page surveying methods (2.1.1.2), and third, corpus-based intra-textbook methods that rely more on quantitative analyses (2.1.1.3).

2.1.1.1 Checklist approach to textbook evaluation

Perhaps the most common approach to evaluating textbooks, and one that will be familiar to many practising teachers, consists in choosing, adapting or developing and then applying checklist-based evaluation frameworks. Typical EFL textbook checklists can feature anything from a dozen (e.g., Garinger 2002) to over a hundred criteria (e.g., Tomlinson et al. 2001). They usually resort to *ad hoc* considerations of the pedagogical and linguistic content of the textbooks, rather than apply any form of empirical measures. Thus, practitioners are expected to be able to answer questions such as: "*Do the exercises and activities in the textbook promote learners' language development?*" (Garinger 2002: 2) or "*Are the grammar rules presented in a logical manner and in increasing order of difficulty?*" (Miekley 2005: n.p.) without resorting to any concrete norms, standards, methodologies or tools.

It goes without saying that attempting to construct a checklist designed to objectively evaluate foreign language textbooks across all dimensions constitutes a truly monumental task. By way of illustration, Tomlinson et al. (2001) devised a set of 133 criteria and used them to each, independently, evaluate eight adult ELF textbooks. The results of their analysis are derived from the mean scores of the four researchers' criteria scores, yet they concede that "the same review, conducted by a different team of reviewers, would almost certainly have produced a different set of results" (Tomlinson et al. 2001: 82). Thus, if checklists are completed without any comparison benchmarks, the results of such checklist-based evaluations risk being largely based on subjective judgement. An advantage of this method, however, is that checklists can easily be adapted to specific teaching contexts. However, this very advantage also entails a risk: rarely are these custom-made checklists thoroughly evaluated in terms of their reliability and validity (Mukundan 2010: 271). For a comprehensive review of checklist-based evaluation frameworks for EFL textbooks, see Mukundan & Ahour (2010).

2.1.1.2 Page-by-page intra-textbook analysis

Before the advent of computer-readable corpora, manual page-by-page surveying of publications was the only way to conduct textbook language studies. In fact, for some types of investigations, this approach is still popular (e.g., Cullen & Kuo 2007; Timmis 2003; Vellenga 2004). By way of illustration, the following section considers the manual intra-textbook methodology applied in Vellenga's (2004) study on pragmatic information featured in EFL grammar textbooks and integrated skills textbooks.

The study is largely qualitative in nature, though Vellenga (2004) provides basic quantitative analysis gleaned from manually counting the number of pages containing pragmatic information (defined as "any information related to culture, context, illocutionary force, politeness, appropriacy and/or register", Vellenga 2004: n. p.) as compared to the total number of pages in each textbook. This page-by-page counting approach is not without its problems and, indeed, Vellenga warns that the resulting "percentages of pages featuring pragmatic information" are somewhat misleading, since, in most cases, pragmatic information only comes in the form of one or two sentences on any one page, thus the page-counting method is prone to producing inflated percentages. In each of the textbooks investigated, Vellenga also counts the number of explicit mentions and metapragmatic descriptions of 21 different speech acts, such as requests, apologies and complaints. Furthermore, she identifies instances of metalanguage in the textbooks[10] and codes them according to four types of functions: description, instruction, introduction, and task-related. Using this data, Vellenga proceeds with descriptive analyses of the types of sentences used in the metalanguage - imperative, declarative and interrogative - and makes *ad hoc* observations about the use of pronouns in metalanguage.

In addition, Vellenga conducted telephone interviews with four experienced EFL/ESL teachers to inquire whether they thought that the textbooks presented issues of politeness and contextual language use in an appropriate manner, and to ask whether the interviewees supplemented textbook materials with additional pragmatic information. Such methodological triangulation can be a very meaningful addition to such an intra-textbook page-by-page analysis but nevertheless bears the same risks observed with the checklist method in terms of poor reliability and validity.

2.1.1.3 Corpus-based intra-textbook analysis

One of the conclusions of Vellenga's (2004) study is that some of the worrying observations in the representations of pragmatic information in EFL textbooks would merit further exploration in a larger study. This could be achieved by replicating the

[10] Note that, here, Vellenga only considered texts "used to preface activities and explain grammatical points" for the analysis of metalanguage, since, as she points out, "[t]he entire contents of a textbook, by its very nature, can be considered metalinguistic" (2004: n. p.).

analysis on a large corpus of EFL textbooks using (partially) automated corpus queries and other corpus-linguistic methods. However, it is questionable as to whether a larger, more quantitative, intra-textbook study would yield results of any greater linguistic or pedagogical significance.

This is illustrated with an example from a corpus-based intra-textbook study assessing the distribution patterns of articles and of their colligation patterns (e.g., *a* + singular count nouns, *the* + ordinal) in a corpus of Malaysian ESL textbooks (Mukundan, Leong Chiew Har & Nimehchisalem 2012). Information on the frequency and distribution of articles and their colligation patterns was extracted automatically from the corpus data. The authors concluded that, in the five textbooks analysed, the article '*an*' is considerably less frequent than the articles '*the*' and '*a*' (see first columns of Table 1). The article subsequently claims that teachers should therefore "create appropriate teaching materials to expose the learners more to the article 'an'" (Mukundan, Leong Chiew Har & Nimehchisalem 2012: 67). However, a quick query of the British National Corpus 1994 (hereafter BNC1994; Burnard 2007) suffices to show that the proportional article frequencies observed in these Malaysian textbooks are, in fact, very comparable to the proportions of article frequencies found in a balanced corpus of naturally occurring English (see Table 1).

Table 1: Distribution of the articles *a*, *an* and *the* in the five Malaysian textbooks examined by Mukundan et al. 2012 (as reported in Table 1, p. 69) and in the BNC1994 (as calculated by Sketch Engine on 13.10.2018) in raw numbers and as a % of articles

	a (*n*)	*a* (%)	*an* (*n*)	*an* (%)	*the* (*n*)	*the* (%)
Textbook 1	1,097	25.88%	141	3.33%	3,001	70.79%
Textbook 2	1,271	32.61%	130	3.34%	2,496	64.05%
Textbook 3	1,630	31.95%	162	3.18%	3,309	64.87%
Textbook 4	1,894	27.90%	209	3.08%	4,685	69.02%
Textbook 5	1,762	25.92%	256	3.77%	4,779	70.31%
Textbook series (1–5)	7,654	28.54%	898	3.35%	18,270	68.12%
BNC1994	2,136,923	25.31%	333,044	3.94%	5,973,437	70.75%

There may well be pedagogical arguments as to why including more explicit teaching material on the article '*an*' may be beneficial but, given that this is far from the case in real-life English usage, textbook authors can hardly be expected to feature all three articles in equal proportions.

Another interesting form of intra-textbook analysis worth mentioning is found in Moreno (2003), in which the accounts of causal metatext (lexico-grammatical features that explicitly signal causal relations) featured in eleven English for Academic

Purposes (EAP) textbooks are compared to the actual expression of causal coherence relations in authentic essays, or extracts of essays, featured in the same textbooks. The author claims to have only included "authentic essays" in these comparisons, but the report does not explain how such an "authentic status" was determined. This is problematic given that it is common for texts featured in textbooks to be presented as "authentic", even if they have been purposefully written as pedagogical material. At the very least, these texts can be expected to have been purposefully chosen to illustrate the linguistic features explained in these textbooks so that their authentic representativeness of general, here Academic, English can be called into question.

2.1.2 Comparative approaches

The previous section explored non-comparative methods to describe Textbook English and, citing Mukundan et al.'s (2012) investigation of articles in Malaysian EFL textbooks and Moreno's (2003) analysis of causal coherence relations in EAP textbooks as examples, pointed to the risks of making pedagogically-motivated evaluations from analyses of textbook language alone. The methods described hereafter involve comparing aspects of the language presented in EFL textbooks with real-life, naturally occurring language data, usually in the form of a reference corpus (2.1.2.3), but also of corpus-based frequency lists (2.1.2.1), or of semi-staged re-enactments of the situations portrayed in the textbook dialogues (2.1.2.5). In particular, the methodologies of two pioneers in the field, Magnus Ljung and Dieter Mindt, will be detailed.

In the mid-1980s and early 1990s, Ljung (e.g., 1990; 1991) conducted an early corpus-driven analysis of the English vocabulary taught in upper secondary EFL classes in Sweden. As part of a large project, his team collected 56 Swedish TEFL publications (designed for the final three years of secondary education) and converted their entire content to machine-readable text. The COBUILD corpus (the main and reserve corpora totalling some 18 million words of mostly written texts; Sinclair et al. 1990) was chosen as a reference corpus, as – at the time – a large collection of contemporary mostly British non-specialist texts. Both corpora were lemmatised and Ljung (1991) subsequently extracted the most frequent 1,000 words in both the pedagogic material corpus and the COBUILD corpus in order to investigate the nature of words unique to either top-frequency word lists, as well as the differences in frequencies between shared words.

Within the confines of this top 1,000-word frequency band, the two corpora shared 796 words. Ljung (1991) analysed the nature of words unique to the TEFL high-frequency list and concluded that the majority of nouns and verbs denote physical objects, processes and human actions, whilst the adjectives express "either emotional judgement (*terrible, wonderful*), physical characteristics (*soft, bright*), or feelings (*angry, glad*)". In contrast, a large proportion of the nouns exclusively found on the

COBUILD high-frequency list denote abstract concepts (*argument, decision, difficulty*), or can be classified as terms from the semantic fields of society or politics (*community, council, campaign, tax*). The high-frequency verbs are predominately used to evaluate human behaviour (*achieve, argue*). Moreover, the majority of the adjectives found in the COBUILD list do not denote physical characteristics (*international, basic, central*).

As for the observed differences in the frequencies of the shared words, in effect, these do little else but reveal that the reference corpus used in these early comparative corpus-based Textbook English studies features mostly elaborate, professionally written and published written texts often on topics quite far removed from those of school textbooks, whereas the pedagogical material appears to have a strong focus on spoken or spoken-like texts. For instance, Ljung (1991) notes that contractions are far more frequent in the TEFL corpus than in the COBUILD. Similarly, he finds that first-, second- and third-person pronouns are more frequent in the TEFL material. Though certainly no mean feat in the 1990s, these results essentially point to the fundamental necessity of drawing on an appropriate reference corpus for the results of such comparative corpus-based textbook analyses to be in any way meaningful.

2.1.2.1 Word-frequency lists approaches

The decades following Ljung's pioneering work, analyses of the vocabulary of EFL textbooks have continued to rely on comparisons of the words found in the EFL publications to corpus-derived frequency word lists. The following section describes more recent and, given modern computing power, less work-intensive, corpus-based methods involving the computation of frequency lists and rates of word repetition in EFL textbooks. The studies chosen to illustrate this approach deal with phrasal verbs (Zarifi & Mukundan 2012) and prepositions (Mukundan & Roslim 2009); however, the method is applicable to any other kind of lexical unit.

Zarifi & Mukundan's (2012) study on phrasal verbs examines a corpus of the spoken sections of five Malaysian secondary school ESL textbooks. First, all occurrences of 19 particles were located using the wordlist function of the software WordSmith (Scott 2011). The researchers then manually identified and tagged the occurrences of phrasal verbs (as opposed to, for example, prepositional uses of these particles). This procedure led to the identification of 108 instances of a total of 66 different *verb + particle* constructions in the spoken textbook corpus. These were then compared to data from the BNC1994 (Burnard 2007).

There are several issues with the presentation and interpretation of the results. First, the quantitative results should be viewed with caution because the textbook corpus explored was relatively small and no statistical testing was carried out so that is it unclear whether many of the observed differences between the textbook and reference

corpora may simply be due to random variation. Second, the interpretation of the results seems somewhat removed from the original pedagogically-orientated aim of the study. For instance, Zarifi & Mukundan (2012: 13) report that 67% of all phrasal verb occurrences in their textbook corpus involve the particles *up*, *down* and *out*. The authors go on to suggest that these results "can be viewed as a deviation of the textbooks from natural use of the language since combination [sic] of 8 particles with 20 lexical verbs has been reported to account for about half of all the combinations in natural language" (Zarifi & Mukundan 2012: 13). However, the study they refer to, Gardner & Davies (2007), reports that the most productive verb particles in British English are, indeed, in the following order: *out, up, back* and *down*. Adding the relevant frequencies presented in Gardner & Davis (2007: 346) reveals that *up, down* and *out* account for 58% of all phrasal verb occurrences in the BNC1994, which is not far off the 67% figure observed in the textbook corpus.

It is also worth bearing in mind that both the token frequencies of phrasal verbs and the number of different types of phrasal verbs varies greatly across different text registers (see, e.g., Liu 2011). It is striking that Zarifi & Mukundan's (2012) textbook corpus only includes textbooks' representations of spoken English, whereas the BNC1994 consists of 90% written registers. Thus, a more meaningful comparison benchmark for this particular Textbook English study may have involved Liu's (2011) study, which reports on the frequency counts of the 150 most frequent phrasal verbs across different registers of the BNC1994 and the Corpus of Contemporary American English (hereafter COCA; Davies 2009; 2010). In fact, on the basis of the detailed results provided in Liu's (2011) appendix, it is possible to calculate that about two-thirds of the most frequently occurring phrasal verbs in spoken registers feature the particles *up*, *down* and *out*.

Other studies comparing frequency lists across a textbook corpus and a reference corpus have sometimes relied on differences in frequency ranks to evaluate the linguistic content of EFL textbooks. Mukundan & Roslim (2009), for instance, discuss supposed differences in the representations of prepositions between Textbook English and naturally occurring English by comparing the frequency ranks of a corpus of ESL textbooks and frequency rank data from the BNC1994. As illustrated in Table 2, the reported rank comparisons do not include the actual frequencies; hence, it is impossible to grasp how large any observed difference in rank actually is. In addition, it is quite reasonable to assume that register differences between the textbooks and the BNC1994 data alone could account for such discrepancies in frequency order. For instance, the high percentage of complex, professionally written texts in the BNC is likely to contribute to a higher frequency of noun phrases and consequently to a more frequent occurrence of the preposition *of* as reported in Table 2.

Table 2: Comparison of the order of the most frequent prepositions in the BNC1994 and three Malaysian ESL textbooks (reproduced from Mukundan & Roslim 2009: 24)

Rank	BNC	Textbooks
1	*of*	*to*
2	*in*	*of*
3	*to*	*in*
4	*on*	*on*
5	*by*	*from*
6	*at*	*at*
7	*from*	*by*
8	*after*	*after*
9	*between*	*before*
10	*under*	*between*
11	*before*	*near*
12	*behind*	*under*
13	*near*	*behind*
14	*in front of*	*in front*

In sum, whilst such corpus-based frequency list comparisons between textbook and reference corpora can produce interesting and pedagogically valuable results, great care must be taken to choose a suitable baseline unit and the appropriate reference corpus, featuring comparable registers – lest the comparison resemble that of apples and pears. Furthermore, providing full quantitative results and/or applying robust statistical testing is essential to the generation of authoritative results.

2.1.2.2 NLP methods

This section attempts to shed light on how corpus-derived word frequency lists may also be used in combination with more complex statistical and natural language processing (NLP) methods to investigate the language of English textbooks. This is illustrated with a method designed to evaluate the development of linguistic complexity across three series of high school Taiwanese ESL textbook series (Chen 2016; 2017).

The method first involves calculating several well-established readability measures for each of the main reading texts of the textbooks. Some of these measures involve phonological analyses of the texts (i.e., they require a tool to identify syllable boundaries) whilst others attempt to account for the complexity of the grammatical

structures (which usually requires part-of-speech tagging or dependency parsing). These measures are combined with an analysis of the vocabulary coverage of the textbook texts. This is computed by extracting the content words from each text and comparing them to a list of the most frequent content words in the BNC1994. Counted from the top of this corpus-based frequency list, the coverage rates of each 1,000-word band is calculated as a percentage. This calculation is repeated for the top thirteen 1,000-word frequency bands, with each percentage representing the proportion of the words in each text that is found in these 1,000-word bands from the BNC1994 list. A variability neighbour-based clustering algorithm is then applied to evaluate the text's complexity on the basis of all of these different measures calculated for each text.

The aim of the method is to tease out relative differences in text difficulty between the textbook texts across the various textbook volumes within each textbook series. Contrary to Chen's (2016) hypothesis that the progression of lexical difficulty ought to be unidirectional – in other words, that both the range of vocabulary and textual structure complexity of the reading texts should increase volume by volume – the results point to some striking non-linear developmental stages of text difficulty across the volumes of the three textbook series.

An advantage of such a method is that it combines a large number of linguistic complexity metrics into one measure that can be used to easily compare the linguistic complexity of texts across different textbooks and textbook series. However, like all of the frequency-list approaches discussed so far, the method requires a suitable reference corpus to produce meaningful results. Furthermore, such methods always involve some arbitrary assumptions as to the appropriate size of the frequency bands employed (i.e., should the analysis focus on the 100, 500, 1,000 or 5,000 most frequent words in any one category?). Hence, no matter how complex, the validity and reliability of such methods remain difficult to ascertain.

Chen (2017) developed an alternative metric, which instead of relying on readability measures, attempts to model the lexical sophistication of textbook texts by examining trigrams, i.e., strings of three consecutive word forms, e.g., *a lot of*. In accordance with his previous research, Chen (2017) concludes that the textbook volumes for advanced learners do not necessarily feature higher degrees of lexical sophistication than previous textbooks in the same series. An advantage of using trigrams to model linguistic complexity is that they also capture valuable information on probabilistic estimates for multiword expressions, thus potentially also revealing relevant developmental trends in the representations of collocation and colligation patterns across textbook texts, volumes, and series.

One methodological issue that remains, however, is that the results of such models are dependent on the number of texts per textbook volume, as well as text length and sentence length. Whilst it may be countered that the latter two are, in fact, desirable features since text length and sentence length may be considered valid factors of lexical and grammatical complexity in their own right, the models nevertheless conflate several variables, thus severely complicating the interpretation of the results that emerge from them. More generally, a major disadvantage of basing Textbook English descriptions and/or evaluation on such complex statistical methods is that the results are highly opaque. If a break in the progression of linguistic complexity across a textbook series is observed, as Chen (2016; 2017) does, it is very difficult to determine which of the many variables entered in the model made consequential contributions to this break in order to understand how improvements could potentially be made. Thus, the pedagogical value of such methods appears relatively limited. Access to such methods is furthermore complicated by the slow uptake of Open Science practices in computational linguistics (see, e.g., Belz et al. 2021; Wieling, Rawee & van Noord 2018). Indeed, it remains relatively common to present new innovative NLP-based methods at conferences and in publications, without ever publishing the corresponding code that would enable (corpus) linguists to apply and assess the methods on new data and research questions.

2.1.2.3 Corpus-based comparisons of 'real-life' language to textbook language

The corpus-based methodologies outlined thus far have relied on corpus-based frequency lists and have thus tended to revolve around the word level. By contrast, the following sections outline comparative corpus-based methodologies applied to the study of Textbook English with the aim of arriving at quantitative and qualitative descriptions of more complex lexico-grammatical features, their functions and pragmatic uses. We begin with Dieter Mindt's pioneering corpus-driven method for the analysis of Textbook English, first described in a monograph on the usage and teaching of future constructions in English (Mindt 1987).

Mindt's interest in textbook language stems from his belief that foreign language textbooks are traditionally based on a pre-conceived grammatical syllabus, rather than on an empirical grammar of actual usage by native speakers (Mindt 1987: 11). He claims that both the grammar syllabus and its content – i.e., the functions of the different grammatical structures – are constructed from two non-empirical, indeed almost anecdotal, sources. He identifies the first as "a longstanding tradition of English language teaching" and the second as "the accepted grammatical knowledge as we find it in current handbooks of English grammar" (Mindt 1997: 40), thus pointing to the cyclical nature of traditional pedagogical grammars. Although his textbook corpus analyses focus on German EFL textbooks, he believes this notion of

a grammatical syllabus to also be true of EFL textbooks published in other countries (Mindt 1997: 40-41).

At the most basic level, the idea behind Mindt's (1987: 9) approach to the analysis of Textbook English is:

> Ein aus Analysen von Sprachkorpora gewonnenes Bild der sprachlichen Realität des heutigen Englisch wird verglichen mit dem Abbild der englischen Sprache, wie es in zwei verbreiteten Lehrwerken dargeboten wird [to compare the results of analyses of authentic English usage that provide a picture of the linguistic reality of present-day English with that of the English language as it is presented in two series of popular [German EFL] textbooks].

As described in 2.1.1.2, page-by-page analysis of textbook language is a difficult, error-prone and time-consuming process. The development of digital data storage and retrieval enabled Mindt to pioneer a new approach to language textbook analysis work using computer-readable textbook corpora. The first step consists in compiling a corpus of naturally occurring, and in Mindt's case, native speaker English. From this data, Mindt extrapolates an empirical grammar of future time expressions that is exclusively based on the observed phenomena, thus breaking with the tradition of introspection-based, deductive grammars. In a third phase, he proceeds with a comparison of the frequencies, functions and co-occurrences of future expressions found in a corpus of two series of popular German EFL textbooks with those from his authentic corpus, the latter thus representing target learner language. Mindt's study of future time expressions in English exemplifies this methodology (1987, 1992).

In this first corpus-driven Textbook English study, Mindt explores the future constructions featured in textbooks designed for lower secondary school (*Sekundarstufe 1, Hauptschule, Klassen 5 - 10*). His reference corpus combines a corpus of English conversation (34 spontaneous recordings of conversations of native British English speaking adults, ca. 170,000 words) and a corpus of contemporary British plays (all published between 1963 and 1980), which he considered to be written representations of natural, spoken language (totalling ca. 184,000 words). Mindt (1987: 50) justifies his choice of a spoken English reference corpus by arguing that the German education authorities stipulate that foreign language teaching at this level should aim to enable students to be able to communicate in everyday situations. His textbook corpus consists of two series of five textbooks each from the two largest publishers of school material in Germany. His analysis focuses exclusively on the language presented in the coursebooks and thus excludes all accompanying material such as workbooks, test material, vocabulary books, and, crucially, the transcripts of the listening exercises (Mindt 1987: 53).

Mindt's approach begins with a comparison of the two reference sub-corpora before comparing these results with those from the analysis of the textbook corpus. Hierarchal and centroid-based cluster analyses are applied to group both reference

sub-corpora. To test the homogeneity of the clusters, i.e., whether differences between the various independent groups are significant, a chi-square test is then applied. Mindt (1987: 62–73) argues that the combination of these two procedures produces a "core description" of the frequencies and co-occurrences of future expressions in spoken British English, which he then goes on to compare to the claims made in major English grammar works and, finally, to the language presented in the two German EFL textbook series. Thus, Mindt's approach involves inferring an empirical grammar inductively, moving from language data to grammatical generalisation, as opposed to the more traditional, deductive approaches that rely on previously ascertained prescriptive rules. Empirical grammars may then be drawn upon to generate pedagogical grammars.[11]

Since Mindt (1987) first exemplified the method by investigating the representations of the future in German EFL textbooks, numerous Textbook English studies have emerged which have, at least partially, been inspired by the Mindtian approach: comparing computer-readable real-life L1 corpora with textbook corpora. So far, these have also focused on specific, individual lexico-grammatical features, such as support verb constructions (Sinclair & Renouf 1988), the indefinite pronouns *any* and *some* (Tesch 1990), modal auxiliaries (Römer 2004b), the progressive aspect (Römer 2005) and *if*-conditionals (Gabrielatos 2013; 2019; Römer 2004a; 2007). Key insights from these studies will be presented in part two (Section 2.2).

2.1.2.4 Comparing textbook language to 'real-life' language

Another approach to exploring the lexico-grammatical content of language textbooks is to manually extract groups of lexico-grammatical features from EFL textbooks and compare these and their frequencies to a reference corpus. This methodology may be perceived as a reversal of Mindt's methodology described in 2.1.2.3. Instead of deriving an empirical grammar of specific features of English from an English L1 corpus to then compare it to the way these features are presented in EFL textbooks, this approach begins with the textbook grammar and attempts to apply it to data extracted from a corpus of naturally occurring English. As this approach, and variations on it, have been applied in many a textbook study, in the following, Koprowski's (2005) analysis of lexical phrases featured in EFL textbooks and Gabrielatos' (2003; 2006; 2013) explorations of textbooks' typologies of conditionals will serve to exemplify the approach.

In an investigation on the usefulness of the lexical phrases presented in contemporary textbooks, Koprowski (2005) manually extracted all the lexical phrases explicitly presented in three intermediate EGP coursebooks. He then compared these 822 lexical

[11] In the following, the term 'pedagogical grammar' will be used according to Dirven (1990: 1) for whom the term 'pedagogical grammar' covers both learning and teaching grammars.

items to data retrieved from five subcorpora of the COBUILD corpus. To this end, a "usefulness score" was calculated for each lexical item extracted from the textbooks. This score relies on two criteria: frequency, arguing that "the commonest units in the language are the ones most likely to be met by learners outside the classroom and should therefore be at the centre of the learning program" (Koprowski 2005: 324) and range, which refers to the number of text types in which a phrase is commonly found, on the grounds that "a unit which exists in a wide variety of registers is generally considered much more useful than an item found in just one, even if that item is highly frequent" (Koprowski 2005: 324). Koprowski's (2005) results are rather disconcerting: 14% of the lexical phrases explicitly featured in the coursebooks were found to be entirely absent from the COBUILD corpus (Sinclair et al. 1990) (see 2.2.1.2 for more details of the results).

'Reversed Mindtian' approaches have also been applied to analyses of Textbook English focusing on grammar and here, too, have pointed to the inadequacy of pedagogical grammars in EFL textbooks. For instance, Gabrielatos (2003; 2006; 2013) examined the typologies of conditional sentences presented in a range of intermediate to advanced EFL textbooks. He identified five types of ELT typologies and concludes that the majority of the coursebooks examined largely follow a simple conditional typology consisting of three types: first conditional with *will*, second and third conditionals with *would.* The most elaborate ELT typologies also include the so-called *zero* conditional, the use of epistemic modals, imperatives, and a range of tenses in all the conditional types. Next, Gabrielatos extracted a random sample of *if*-sentences from the written sections of the BNC1994 and annotated them according to their conditionality, tense and aspect marking, time reference, modality, etc. Using these annotations, he calculated how many of these naturally occurring *if*-sentences can accurately be described according to the ELT conditional typologies presented in EFL textbooks. Using the most basic typology, only 15% of the *if*-sentences from the BNC could be successfully classified. Most strikingly, even with the most complex and inclusive of the ELT typologies identified in the textbooks, 22% of all sentences are still unaccounted for. These results lead Gabrielatos (2006: 2) to conclude that the typology explicitly taught in textbooks "provides learners with an incomplete, and in some cases distorted, picture of *if*-conditionals".

Contrary to Mindt's studies, Gabrielatos' method only explores ELT grammar as presented in the textbooks' grammar sections, rather than, more holistically, to the totality of the language to which learners are exposed via their textbooks. Methodologically, such 'reversed Mindtian' approaches have the disadvantage of only pointing to the inadequacy of Textbook English without providing textbook authors and EFL teachers with those "useful" (to keep with Koprowski's terminology) linguistic features that are missing from or grossly underrepresented in textbooks but which corpus-based studies have demonstrated to be highly frequent and salient.

Though a much more time-consuming undertaking, the compilation of a textbook corpus, in addition to the qualitative analysis of their grammar sections, allows for the analysis of how specific linguistic features are represented in the textbooks' reading passages, exercises, instructions and listening transcripts. Following such an approach, Winter & Le Foll (forthcoming) revealed that a remarkably large proportion (between 43% and 53% depending on the typology applied) of *if*-conditionals featured in 42 secondary school EFL textbooks did not fit the conditional typologies presented in these same textbooks; thus highlighting a significant gap between what is taught in textbooks and what is practiced by textbook authors within these same textbooks (on conditionals in EFL/ESL textbooks, see also Gabrielatos 2006; 2013; Römer 2004a; 2007; Tesch 1990).

2.1.2.5 Elicitation approaches

The previous section explained how lexical units or grammar rules extracted from EFL textbooks can be compared to reference corpus data to form judgements as to how "useful" (e.g., Koprowski 2005) or "accurate" (e.g., Gabrielatos 2003; 2006; 2019) textbook input is likely to be for foreign language learners. In the following, a different approach to evaluating the authenticity of textbook language will be described. It relies on the re-creation of communicative situations simulated in the textbooks. In this respect, it can be said to share some characteristics with the 'reversed Mindtian' approach described in the previous section.

In a study of spoken discourse features in Textbook English, Gilmore (2004) investigates the authenticity of the language presented in service encounter dialogues (e.g. hiring a car from a rental shop, or asking for directions in the street) in ten EFL textbooks published between 1981 and 2001. To this end, he selected one such dialogue from each textbook and extracted, in note form, all the questions asked by the information receiver in each dialogue. The questions were then reformulated and used as a basis for real conversations in the genuine settings imagined by the textbook authors. The real dialogues were recorded and transcribed and subsequently used to compare the use of discourse features in the textbook dialogues and their 'authentic' re-creations.

Gilmore's (2004) method is highly original, yet it appears difficult to draw conclusions on results based on just seven textbook dialogues and seven enactments. Critically, the authentic nature of semi-staged service encounter dialogues (and, though this is not specified in the publication, perhaps even with the researcher acting as the information receiver in each of the re-created dialogues) may be questioned. Furthermore, it would be interesting to investigate to what extent discourse features of the conversations would differ if the information receiver were a non-English native speaker, since the majority of these service encounter dialogues are intended to

present typical communicative situations that tourists may face in an English-speaking country.

In a similar vein, Schauer & Adolphs (2006) explore the possibility of using native speakers' responses to discourse completion tasks, rather than large-scale native speaker corpora to inform the teaching of formulaic sequences in the EFL classroom. Expressions of gratitude featured in four EFL textbooks were compared to those elicited in discourse completion tasks, as well as to thanking formulae retrieved from the spoken CANCODE corpus. Unsurprisingly, the researchers observe notable differences between the controlled, elicited responses and the natural conversations found in the corpus. They argue that the first type of data can facilitate the acquisition of more recent language pattern changes, whilst the latter can generally provide "[a] much broader picture" such as "insights into the procedural aspects of expressing gratitude", which may materialise in the form of collaborative negotiation or re-lexicalisation of another speaker's utterance (Schauer & Adolphs 2006: 130).

In sum, it is tempting to conclude that constructed dialogues, whether in the form of semi-staged re-enactments of textbook dialogues or discourse completion tasks, are unlikely to yield sufficiently robust data to reliably evaluate the authenticity of textbook language. Nonetheless, such methodological approaches can point towards aspects of textbooks that may need to be updated or re-organised. They can thus provide valuable starting points for further investigations.

2.1.2.6 Adding learner corpora to the equation

A number of studies explore Textbook English with a view of better understanding learners' interlanguage. Since textbooks constitute a major source of secondary school learners' L2 input (see 1.6), it may be speculated that learners' over-, underuse or misuse[12] of particular lexico-grammatical features may be (at least partly) attributed to their textbooks' treatment of these features. In order to investigate such potentially causal relationships, some studies have attempted to triangulate results derived from textbook vs. reference corpora comparisons with insights from learner corpora (e.g., Fujimoto 2017; Gabrielatos 2013; Möller 2020; Rankin 2010; Vine 2013; Winter & Le Foll forthcoming). The potential and limitations of such methodologies are exemplified in the following.

An example of a study drawing on textbook data to glean insights into English learners' difficulties is Rankin's (2010) study of adverb placement in L2 essay writing. In this study, 37 English essays written by Austrian university students were surveyed

[12] Note that, unlike 'misuse', the terms 'overuse' and 'underuse' are descriptive, not prescriptive, terms; they merely refer to the fact that a linguistic form is found significantly more or less in the learner corpus than in the reference corpus.

for adverb placement errors. The errors were checked against the Louvain Corpus of Native English Essays (LOCNESS; Granger 1998) and tagged. In parallel, all the pedagogic material used during the students' English language course (the duration of which is not mentioned) was gathered; materials and exercises specifically dedicated to adverbs were tallied. In other words, Rankin's study exclusively looks at explicit practice of adverbial usage, rather than at all the lexico-grammatical constructions involving adverbs to which learners were exposed in class. In the qualitative part of the analysis, Rankin (2010) compares the students' adverb placement errors with the classroom input material. He concludes that whilst the adverb grammar exercises provided often require learners to choose appropriate adverbs for particular gaps in gap-filling exercises, they do little to address the issue of adverb placement within sentences. He stresses that "residual problems with adverb placement are not due to any major deficiencies in basic grammar but rather to the fact that appropriate variation in adverb placement for special discourse and pragmatic contexts has not been mastered" (Rankin 2010: 214).

In another example of an English Textbook study involving learner corpus data, Fujimoto (2017) examines the use of the present perfect simple with and without temporal adverbials across three corpora: a longitudinal learner corpus of Japanese university students' academic writing assignments, a textbook corpus consisting of "reading passages" drawn from six high school English textbooks and a reference corpus consisting of the fiction and general prose subcorpora of the FLOB (representing British English of the early 1990s; Hundt, Sand & Siemund 1998) and Frown (representing American English of the early 1990s; Hundt, Sand & Skandera 1999) corpora. The frequencies of the co-occurrences of the present perfect simple with temporal adverbials in the reference corpora are compared to the corresponding frequencies extracted from both the learner corpus and the textbook corpus. As Fujimoto demonstrates, such L1 vs. textbook vs. learner corpora comparisons can provide relevant insights into the source of learners' difficulties with regards to specific lexico-grammatical features. However, in this case, it may be argued that the FLOB and Frown fiction and general prose subcorpora are questionable baselines for comparisons of reading texts from secondary school textbooks and student academic writing. Academic writing is known to follow quite typical register-specific lexico-grammatical patterns and the study fails to account for such register discrepancies.

Following a similar procedure, Vine (2013) computed the frequency of four high-frequency category ambiguous words (*down, like, round* and *up*) across English native language (hereafter ENL) corpora (spoken and written British English and New Zealand English), learner English corpora (spoken and written) and an EFL textbook corpus. Comparisons of the frequencies of occurrences of each of these four words sorted in terms of the grammatical category of each use revealed considerable variations across all the corpora. It is interesting to note that whilst Vine subdivides

all of her results into the spoken and written subcorpora for the ENL and the learner English corpora, this register differentiation is not made for the textbook corpus, even though language textbooks typically include registers as diverse as conversation, newspaper writing and fiction (see 3.3.1.4). Here, too, it is difficult to draw any meaningful conclusions from the results of this analysis since the frequencies reported for each part-of-speech use also vary greatly across the different reference subcorpora. Nevertheless, it would appear that the frequencies of the learner corpora are considerably closer to those of the textbook corpus than to those observed in the reference ENL corpora. Such observations lend tentative support to the hypothesis that, given they represent a major source of L2 input, textbooks play a crucial role in EFL learners' language acquisition processes (see 1.6–1.7) but also remind us that mode and register differences need to be accounted for when describing and evaluating the language of EFL textbooks.

2.1.2.7 Textbook language as learner target language

As in the previous section, this section presents a methodological approach to the study of Textbook English that also compares data from a learner corpus to that of a textbook corpus. However, as illustrated with a study by Tono (2004), this particular approach not only adds a layer of cross-linguistic comparison using an L1 (Japanese) corpus, it also turns the equation around by assuming textbook language to be the learners' target language.

In stark contrast to the corpus-based comparative approaches reviewed so far, Yukio Tono (2004: 51) claims that "textbook English is a useful target corpus to use in the study of learner language". He convincingly argues that comparing learner language to texts produced by native speaker professionals makes little sense. Indeed, all of the well-known general English corpora used in most of the Textbook English studies reviewed so far (e.g., BNC1994, Brown, FLOB, etc.) predominantly feature professionally written or spoken texts such as newspaper articles, extracts of novels and political speeches. Whilst Tono recognises that the use of such reference corpora may make sense when it comes to supporting advanced L2 learners or professional translators, he considers that the majority of English learners in Japan have no such aspirations:

> In the present case, it is certainly not the language of the BNC that the Japanese learners of English are aiming at, but, rather, a modified English which represents what they are more exposed to in EFL settings in Japan (Tono 2004: 51).

Although he acknowledges that Textbook English often does not reflect actual language use, he nevertheless argues that, since Textbook English is constructed so as to facilitate learning, it makes sense to apply Textbook English as a benchmark when investigating EFL attainment.

In the Japanese context, Tono (2004) emphasises the fact that textbooks represent the primary source of English language input, noting that even when teachers use English as medium of instruction, they tend to restrict themselves to the structures represented in the textbooks. Hence, Tono (2004: 52) claims that "it is fair to say that the English used in ELT textbooks is the target for most learners of English in Japan". Whilst the present author disagrees with the idea that Textbook English necessarily is (see also Timmis 2003) or should be the learners' target (see 2.1.2.7), Tono undoubtedly raises an important point: the need to reflect on the suitability of using general English corpora such as the BNC as benchmark reference corpus when analysing both textbook and learner language (see Winter & Le Foll forthcoming for an example of a study that justifies the use of only specific subcorpora of the BNC in such a comparative analysis).

In a study on the acquisition of English argument structures by Japanese learners, Tono (2004) compares three different types of corpora: a) an interlanguage corpus of free compositions written by Japanese learners of English, b) a native language corpus consisting of English newspaper articles, and c) an EFL textbook corpus. First, sentence frame patterns with the three most frequent verbs (except *have* and *be*) are extracted from all three corpora. For each high-frequency verb use, Tono collects data on a range of variables, e.g., frequency in the textbook corpus, number of learner errors, learners' year group, Japanese equivalents of the verb constructions, etc. Log-linear analysis is used to tease out the most important factors influencing Japanese learners use of these sentence frame patterns. To this end, all frequencies are converted to categorical data (i.e., to high, mid, or low occurrence); thus, presumably reducing the degree of accurateness and adding a layer of arbitrariness in the statistical analysis. The results of the best fitting models show that the learners' school year exerts the most influence on learners' idiomatic production of sentence frame patterns. Interestingly, the second most influential factor is the frequency of a pattern in the textbook corpus. Strong two-way interaction effects between the factors 'school year' and 'textbook frequency' are also observed. By contrast, 'learner error' only significantly interacts with 'school year in one case (for the verb *get*). This suggests that textbook frequencies mostly impact students' overuse or underuse of a particular verb pattern, rather than their rate of success in producing the pattern idiomatically. In addition, the results also show that whether or not a verb argument structure has a comparable equivalent structure in Japanese has less impact on Japanese L1 learners' production of the target structures than how often the structure is featured in the textbooks they learn from.

Although working from radically different premises, both this section and the one preceding it have revealed the value of integrating learner corpora in Textbook English evaluation. At least since the late 1990s, a number of academics have advocated integrating observations gleaned from learner corpora into the design of

new EFL publications (Granger 1998; e.g., Kaszubski 1998). Indeed, some major textbook publishers have now latched onto the idea; as mentioned in 1.8, Cambridge University Press now draws on the error-tagged Cambridge Learner Corpus, which was compiled on the basis of student responses taken from Cambridge English Language Assessment examinations.

2.1.3 Evaluating the impact of textbook language

Most of the studies of Textbook English outlined thus far have aimed to describe the linguistic input of EFL textbooks. By contrast, this section examines studies that also aim to evaluate the potential pedagogical impact of this textbook-based input. Of course, language teachers regularly reflect on the quality of the textbook materials they introduce in class and will thus periodically conduct at least impressionistic retrospective analyses of the textbooks' content. However, attempts to formalise and quantify such retrospective evaluations on the effectiveness of foreign language textbooks have tended to focus on the nature of the tasks and activities featured in the textbooks (see, e.g., Ellis 1997), rather than home in on the quality and usefulness of their linguistic input. This section presents two studies that investigate the linguistic content of EFL publications with regards to their impact on learners in terms of learning outcomes and efficacy. As an extension of comparative corpus-based approaches described in 2.1.2.3, the methodology of the first study (Alejo González et al., 2010) will be familiar to the reader. The methodology of the second study (Gouverneur 2008a), however, relies on the analysis of a corpus of textbook activities annotated with a complex pedagogical annotation scheme.

Alejo González et al. (2010) delve into both the implicit and explicit mentions of phrasal verbs in textbooks, focusing on the learning efficacy gains for the textbook users. To this end, they select eight popular EFL textbooks targeted at the Spanish secondary school market. Their research on the likelihood of incidental learning of phrasal verbs in the ELT material is based on frequency counts within the textbooks and on frequency comparisons with the BNC1994. They report that the vast majority of the phrasal verbs featured in the examined textbooks only appear once or twice in any one textbook, and thus do not occur nearly frequently enough to warrant incidental learning. Moreover, comparisons of the frequencies of the 25 most frequently occurring phrasal verbs from the BNC1994 with the data from the textbooks show that while two of those phrasal verbs (*go out* and *look after*) are vastly over-represented in the ELT material, many others are largely under-represented (e.g., *go back*, *point out* and *take over*), if not entirely absent (*carry on*).

The explicit part of the investigation examines the metalanguage used to describe phrasal verbs and related phenomena in the textbooks, as well as the types of exercises designed to encourage the acquisition of these lexical items. Referring to pedagogical approaches inspired by cognitive linguistics (see 1.4), Alejo González et al. (2010)

deplore that none of the textbooks examined organise explicit mentions of phrasal verbs in a way that is likely to facilitate acquisition by encouraging learners to understand the 'motivated' nature of the particles in combination with their corresponding lexical verbs (for more on the cognitive linguistics' view that phrasal verb particles display a certain degree of compositionally, see, e.g., Condon 2008; Spring 2018; Torres-Martínez 2019; Tsaroucha 2018). Alejo González et al. (2010: 72) argue that "[i]f materials create too few opportunities for incidental uptake, then this should be compensated by explicit targeting" and conclude that their sample of eight Spanish secondary school EFL textbooks fail to adequately do so.

In conclusion to a large-scale learner corpus study on collocation, Nesselhauf (2005: 238) also postulates that collocations are not taught in a way that spurs on their idiomatic acquisition since there appears to be no correlation between the number of years of classroom teaching and the idiomaticity of the collocations learners produce (see also Le Foll 2016). Although Gouverneur (2008) does not directly compare Textbook English to learner language, she takes Nesselhauf's (2005) corpus-based learner error analysis as her starting point and designs a study that aims to tease out whether "learners' deficiencies in the production of phraseological patterns of simple verbs might be teaching-induced or, more precisely, material-induced" (Gouverneur 2008a: 224). To do so, Gouverneur (2008) draws on a textbook corpus (TeMa; Meunier & Gouverneur 2009) which includes the full pedagogical materials from each textbook series including the reading texts, transcripts, vocabulary exercises and instructions from both student's coursebooks and workbooks. Uniquely, the TeMa corpus also includes detailed pedagogical annotation of the subcorpora containing the vocabulary exercises with some 80 codes referring to various aspects of task design and content (Meunier & Gouverneur 2009). As part of this study, all the instances of *make* and *take* were automatically retrieved from the vocabulary exercise subcorpora and the results were manually sorted for meaning and collocational patterns. High-frequency verbs are found to feature prominently in the context of restricted collocations in all the textbooks, thus suggesting that material designers had taken due care "to include a significant number of phraseological uses [of *make* and *take*] in the exercises" (Gouverneur 2008: 234).

Next, all instances of restricted collocations identified were categorised according to the degree of focus on the collocation in the corresponding exercises. It transpires that direct, explicit focus on these lexical units was largely found in the intermediate level textbooks. Gouverneur (2008: 235) notes that, in more advanced textbooks, these collocations are no longer dealt with explicitly. This trend was found to be true for all three series of textbooks examined. Gouverneur (2008: 235) suggests that:

> [t]his lack of direct focus on restricted collocations at the advanced level might well be one of the reasons why more proficient learners have so many problems dealing with high-frequency verbs.

The vocabulary exercises were also annotated according to eight types of pedagogical activities, which were themselves grouped into four larger categories corresponding to the cognitive processes they are (presumably) designed to activate. According to this annotation scheme, whilst 12% of the intermediate learning activities on collocations of *make* and *take* are designed to activate understanding, such activities are entirely absent from the advanced textbooks (Gouverneur 2008a: 236–237). Another striking finding is that fewer than 20% of all the advanced exercises require learners to produce an answer that requires full retrieval from the mental lexicon. Most exercises merely require students to select the correct solution from a given list of words or expressions (Gouverneur 2008a: 236–237).

2.2 Part Two: Key findings of Textbook English studies

Part one of this literature review chapter provided an overview of the wide range of methodologies that have so far been applied to survey the linguistic content of EFL textbooks, reporting on the results of individual studies to illustrate the advantages and potential weaknesses of the various methods. Part two, by contrast, homes in on some of the key results of previous studies examining the language of English textbooks. Whenever possible, emphasis is placed on the results of EFL textbooks designed for secondary school contexts but, where studies are sparse and they are deemed to be relevant, the results of adult EFL, EGP, ESL and EAP textbook studies are also mentioned.

The following section falls into three subsections. First, the results of studies principally exploring the lexis of Textbook English will be summarised. Second, studies investigating more complex lexico-grammatical features denoting verb tense, aspect and argument structures are presented before, third, the results of the few Textbook English studies focusing on pragmatics and discourse are reviewed. Note that this division of the examined linguistic phenomena into these broad categories purely serves an organisational purpose. Indeed, and as will be made evident in the discussion of the studies' results, many of the examined linguistic features straddle any artificial boundaries between lexis, grammar, discourse, semantics, and pragmatics.

2.2.1 Lexis

Perhaps the most immediately obvious aspect of Textbook English is its vocabulary – in other words, the range of words and multi-word units presented in English textbooks. In the following, the results of a small selection of studies focusing on the lexis of English textbooks are outlined. Subsection 2.2.1.1 focuses on individual words whilst 2.2.1.2 looks at the treatment of multi-words units such as collocations, phrasal verbs and lexical bundles in Textbook English.

2.2.1.1 Individual words

The tradition of examining the vocabulary of EFL textbooks goes back a long way. The results of Ljung's (1990; 1991) analysis of the vocabulary featured in upper secondary school Swedish EFL publications have already been discussed in 2.1.2.1. As a reminder, the studies pointed to an overrepresentation of concrete words to the detriment of abstract ones and deplored the poor representation of lexical units commonly used in communicative interaction and in the establishment of social relationships.

As part of another early corpus-based Textbook English study, Renouf (1984; cited in Sinclair & Renouf 1988) investigated learners' vocabulary input in nine major EFL coursebooks. Her analysis shows that, in the first coursebook of each series, the number of different word forms introduced ranged from just over 1,000 to nearly 4,000 – thus representing an incredibly wide variation. The average rate of re-occurrence of each word form across the different textbook series was also calculated. Here, too, the patterns of reinforcement also ranged widely: from six to 17 times.

Based on an analysis of the same textbook series, Sinclair & Renouf (1988) explored learners' exposure to delexical constructions (also frequently referred to as support verb constructions). The authors concluded that such constructions are mostly neglected in Textbook English, in spite of their preponderance in ENL corpora. This study, however, disregards occurrences of delexical constructions occurring within "the rubric of the text" as of secondary importance, rather than as an integral part of the teaching programme (Sinclair & Renouf 1988: 153). Thus, whilst Sinclair & Renouf (1988) deplore that ditransitive uses of the verb *give* are not explicitly highlighted in the coursebooks, they acknowledge that such patterns are featured within the coursebooks' "text rubrics".

A few decades later, Reda (2003) conducted a large-scale analysis of (adult) EGP textbooks designed for the global EFL market. The study concludes that the vast majority of textbooks across all proficiency levels are largely based on a "limited number of 'general interest' topics", such as *cooking, food and drink* or *holidays and travel* (Reda 2003: 264). Hence, in spite of the rise of English as an international language in the context of globalisation, the lexical syllabus taught in the EFL/EIL textbooks examined confines itself to "the basic area of the English vocabulary – the 'visitors' wing'" (Reda 2003: 268). Even the more advanced coursebooks in each series do not depart from these "basic topics" of "general interest".

Whilst Reda's (2003) analysis of English textbooks targeted at adults appears to point to a common understanding by textbook publishers as to the "topics of general interest" to be covered in EFL textbooks, Catalán & Francisco (2008) conclude that the textbook authors of EFL textbooks used at two levels in Spain (6th grade of

primary education and 4th grade of secondary education) disagree on the core vocabulary learners ought to acquire at these stages. The authors measure the number of tokens and types for each textbook and compare type-token ratios. Moreover, they compute lists of the 50 most frequent content words from each textbook. Comparisons of these frequency lists show that Spanish learners of English are exposed to very different words and with varying frequencies depending on which textbook they have been assigned. Catalán & Francisco (2008: 161–162) point out that the Spanish authorities do not specify how many or which words students ought to have acquired by any particular stage and conclude from their study that the textbook authors also appear to lack a systematic approach to vocabulary selection and presentation.

Whilst the studies reviewed so far have focused on the breadth of vocabulary covered by English textbooks, the following studies examine three specific functional categories of words: linking adverbials (Conrad 2004), the definite article (Yoo 2009), and adjectives (Biber & Reppen 2002a).

Conrad (2004) focuses on the frequencies and usage of linking adverbials of contrast and concession in two registers (conversation and academic prose), comparing data from the *Longman Grammar of Spoken and Written Language* (Biber et al. 1999) to the coverage of the adverbial *though* in four American ESL textbooks. The study concludes that textbook coverage does not match native corpus evidence. For instance, Conrad (2004) notes that three out of the four textbooks fail to include the use of *though* as a linking adverbial, and that the only textbook that mentions it presents it as a means of showing contrast but neglects its usage as a means to express concession. Although it occurs frequently in L1 conversation, all four textbooks fail to mention *though* as a means of softening disagreement between speakers. Conrad (2004) observes that only one textbook suggests a number of contrast linking adverbials to use in conversation, but that this textbook misleads learners into thinking that *however* and *on the other hand* are commonly used in conversation, whereas they, in fact, occur far more frequently in academic prose than in any other register. Indeed, a number of Textbook English studies have pointed to the predominance of lexico-grammatical features typical of written registers in textbook dialogues designed to emulate spontaneous spoken interaction (see 2.2.4 on spoken grammar in Textbook English).

In a study following a very similar approach, Yoo (2009) compared the treatment of definite articles in six EFL/ESL grammars with corpus findings reported in the *Longman Grammar of Spoken and Written English* (Biber et. al 1999). The results suggest that whilst most ESL/EFL grammars extensively describe the anaphoric and associative uses of the definite article (e.g., *Let's go to the Indian restaurant. <u>The</u> food is delicious.*), its situational (e.g., *Can I have <u>the</u> chutney, please?*) and cataphoric uses (e.g., *At <u>the</u> beginning of my PhD*) are neglected. The findings

potentially have important pedagogical implications since corpus data shows that the situation and cataphoric uses of the definite article are more common than its anaphoric use in a number of text registers that English learners are highly likely to be confronted with: namely, conversation, newspaper language and academic prose (Yoo 2009: 273–276).

In sum, the results of these case studies on the presentation of *though* and *the* in textbook grammars serve as a reminder as to the central importance of production modes, registers, and text types in the (contextual) use of specific lexico-grammatical features. It therefore follows that these factors must be taken into consideration, both in the elaboration and evaluation of Textbook English (see 3.1). The final case study on individual words reviewed as part of this section, Biber & Reppen (2002), illustrates how the proficiency levels that textbooks are targeted at must also be considered when modelling Textbook English.

Among other lexico-grammatical phenomena, Biber & Reppen (2002) focus on the role of adjectives in Textbook English. To this end, they compare the frequencies of different types of nominal premodifiers in a large general English corpus with how they are presented in six popular ESL/EFL grammar textbooks. The results suggest that the pedagogical materials over-emphasise the prevalence of participial adjectives (e.g., *an exciting game, an interested couple*) whilst underestimating the pervasiveness of nominal premodifiers (e.g., *a grammar lesson*) (Biber & Reppen 2002a: 201–202). As far as teaching beginner-level conversation is concerned, a focus on attributive adjectives (e.g., *the big house*) appears to be justified. At higher levels of proficiency, however, the authors argue that students would likely benefit from greater exposure to the use of nouns as nominal premodifiers since corpus-based findings have shown that these are conspicuously frequent in both newspaper and academic writing (Biber & Reppen 2002a: 202). Thus, these results point to the necessity of not only accounting for mode and/or register differences, as highlighted in the discussion of Conrad's (2004) and Yoo's (2009) results, but also textbook proficiency levels when describing and evaluating the language of textbooks. However, the vast majority of Textbook English studies reviewed as part of this chapter do not account for either of these potential sources of variation.

2.2.1.2 Multi-word units

Though 2.1.2.1 has shown that the study's methodology is not without its flaws, Zarifi & Munkundan (2012) certainly point to a disconcerting gap between the phrasal verbs featured in the 'spoken' sections of Malaysian ESL textbooks and the most frequent phrasal verbs in the BNC1994. For instance, they report that the most frequently occurring phrasal verbs in their textbook corpus, *clean up* and *melt down*, do not, in fact, belong to the most frequent 100 phrasal verbs in the BNC1994. The results also reveal that other highly frequent and more pedagogically valuable phrasal verbs –

such as *work out, turn over* and *go over* – do not appear at all in this corpus of textbook dialogues (Zarifi & Mukundan 2012: 13).

The comparative corpus-based methodology employed in Koprowski's (2005) exploration of multi-word lexical units in Textbook English was laid out in 2.1.2.4. Among the most striking results was the fact that more than 14% of the lexical phrases explicitly featured in the three EGP coursebooks examined were entirely absent from the selected reference corpus, the COBUILD (Sinclair et al. 1990). Such phrases include *cheap steak*, *mild cigarette*, *imprisoned man*, *recommend fully* and *on its last feet* (Mark Koprowski 2005: 328). Based on this small-scale investigation, Koprowski (2005: 329) draws the provocative conclusion that "the more lexical phrases in a course[book], the less useful the items tend to be on average". It would appear that textbook authors frequently attempt to supply excessively comprehensive sets of lexical phrases of a single type or on a single topic thus resulting in the inclusion of some highly infrequent, sometimes outright implausible, collocations. In addition, and just like Catalán & Francisco's (2008) study of individual words in school EFL textbooks, Koprowski (2005) points to a striking lack of consensus as to what constitutes a meaningful lexical curriculum at intermediate level since less than 1% of the lexical phrases collected are shared by any of the textbooks under study (Mark Koprowski 2005: 330).

Another strand of Textbook English studies is concerned with phrasemes, i.e., recurrent sequences of words such as *the fact that, I want you to* and *which is why* (also referred to as lexical bundles, lexical clusters and n-grams). Even though they often do not represent a complete structure nor are they necessarily idiomatic in meaning, these multi-word units nevertheless capture important discourse functions in both written and spoken registers (Biber 2006: 134–135; Biber & Barbieri 2007: 264) and are thus very relevant to the description and evaluation of Textbook English.

Siepmann (2014) compared the phrasemes featured in the vocabulary sections of two series of German secondary school EFL textbooks with a revised version of Martinez & Schmitt's (2012) list of the most frequent "non-transparent phrasemes" found in the BNC1994. Across these entire textbook series spanning five years of EFL instruction, only 12% (for *Green Line)* and 16% (for *G21*) of the phrasemes of the revised corpus-based list were mentioned at least once. Siepmann (2014) concludes that the selection of phrasemes in these textbooks is seemingly not based on frequency or, in fact, on any other systematic set of criteria. In addition, and contrary to expectations, it was also not the case that the number of phrasemes featured in these textbooks rose as students' proficiency level increased.

Aside from the aforementioned study, most Textbook English research to have taken a phraseme or lexical bundle perspective have examined English for Academic

Purposes (EAP) and English for Specific Purposes (ESP) textbooks. These materials are designed to equip non-native speakers of English with the necessary skills to cope with the demands of academic reading and writing at English-speaking universities and/or (future) professional activities in English. A number of studies have attempted to describe and/or evaluate the language of such textbooks by examining the types and frequencies of lexical bundles they feature (e.g., Biber et al. 2002; Biber 2006; Chen 2010; Grabowski 2015; Wood 2010; Wood & Appel 2014).

Wood (2010), for instance, investigates the frequency of lexical clusters in six intermediate and advanced EAP textbooks. This corpus-based analysis of the textbook materials reveals that textbook instructions feature considerably more lexical clusters than the reading passages. Wood (2010) advances the theory that publishers aim for a certain amount of consistency in the formulation of the tasks, thus leading to a high frequency of lexical clusters in the instructional texts. The reading passages, on the other hand, contain fewer lexical clusters and their frequencies of occurrence within any one textbook are such that the author believes that it is unlikely that learners can acquire them solely through reading. Wood's (2010) page-by-page analysis of the pedagogical treatment of formulaic language in the examined textbooks strengthens this hypothesis, as no attempt appears to be made to focus learners' attention on these lexical units.

The presentation of lexical bundles in EAP and ESP textbooks is also the focus of Wood & Appel's study (2014) in which they first extracted the most frequent three and four-word bundles in a corpus of ten first-year business and engineering textbooks, and then queried a corpus of five intermediate and advanced EAP textbooks to reveal which of those bundles appear in the EAP textbooks. Depending on the EAP textbook, between 35% to 47% of the most frequently occurring lexical bundles from the subject textbooks were found at least once in the EAP textbooks. However, the authors deplore that none of the formulaic sequences are dealt with pedagogically, i.e., presented as units worth learning or highlighted in any way that might raise learners' awareness of their potential.

Focusing on one discipline only, electrical engineering, Chen (2010) also compares the frequency and nature of multi-word units in entry-level university electrical engineering textbooks and ESP textbooks especially designed for students of this same discipline. In contrast to Wood & Appel (2014), however, she not only compiles a list of the most frequent lexical bundles found in the introductory subject-specific textbooks, but also one for the ESP textbooks. As a result, she is able to compare the types, frequencies and pragmatic functions of lexical bundles featured in both types of textbooks. Her results match those of Wood & Appel (2014) in that only a third of the lexical bundles identified in the electrical engineering introductory textbooks occur at least once in the corresponding ESP textbooks. Furthermore, a qualitative

analysis of the pragmatic functions of the bundles demonstrates that entire subcategories of stance bundles (e.g., *can be used to*, *it is important to*) are missing from the ESP textbooks. When it comes to referential bundles (e.g., *is referred to as*, *a great deal of*), Chen concludes that "the ESP textbooks underrepresent quantity and spatial specifications but overemphasize referential information which is not central in target language use, such as the introduction to new concepts/definitions and provision of time information" (Chen 2010: 123).

2.2.2 Tense and aspect

Having presented some of the observations derived from a range of studies on the lexis of Textbook English, this section now turns to the representation of tenses and aspects in EFL publications. To this end, it seems natural to begin with one of the earliest comparative corpus-based studies already mentioned in 2.1.2.3: Mindt's (1987; 1992) study on the prevalence, functions and lexico-grammatical patterning of future constructions in German EFL textbooks.

2.2.2.1 Future constructions

As explained in 2.1.2.3, Mindt (1987; 1992) undertook to study the future constructions presented in textbooks designed for the first cycle of German secondary schools with those actually produced in speech by British native speakers. He concluded that the examined German EFL textbooks under-represent *will*, over-represent *going to* and leave out *shall* as a means of expressing future situations altogether (Mindt 1992: 189). Mindt (1992: 37) also observed that, compared to his reference corpus of (pseudo-)spoken English, the contracted forms of *going to* were considerably under-represented in the spoken passages of the textbooks. Furthermore, he interprets the absence of *gonna* and *ain't* in the examined textbooks as a misrepresentation of English language usage at the time of the study (Mindt 1992: 35, 41). Mindt's analyses go beyond simple comparisons of relative frequencies and also explore the context in which certain expressions of the future co-occur. For instance, he notes that *going to* co-occurs with a relatively narrow range of expressions in Textbook English and hypothesises that this is due to an over-generalisation of *going to* as a structure used almost exclusively to express the future (Mindt 1992: 190).

2.2.2.2 The present perfect

Following a similar approach, Schlüter (2002) conducted a book-length corpus-driven analysis of the use of the present perfect on the basis of a native-speaker corpus consisting of spoken and written English. From this data, Schlüter (2002) established a so-called 'empirical grammar' (see also Dirven 1990; Mindt 1995a) of the present perfect and contrasted it to existing traditional grammars that are known to mostly rely on introspection, as well as to the grammar sections of two popular series of

secondary school EFL textbooks used in Germany, together with their accompanying grammar and activity books. The textbooks are found to present the functions of the present perfect in substantially different ways (Schlüter 2002: 219–328). For example, textbooks fail to explain that the present perfect progressive is often used to refer to iterative actions or events rather than continuous ones.

In order to investigate Japanese EFL learners' difficulties with the use of the present perfect, Fujimoto (2017) triangulated results from three corpora: a reference corpus of general American and British English, a Japanese L1 English L2 learner corpus and a corpus of EFL textbooks designed for Japanese high schools. Fujimoto (2017) reports that learners overuse the present perfect with temporal adverbials as compared to the reference corpus. The six textbooks examined vary greatly in their use of the simple present perfect both with and without temporal adverbials. However, the two textbooks that radically over-represent the simple present perfect with temporal adverbials, as opposed to without, do so principally in the exercises, rather than in the extended reading passages. Fujimoto (2017) suggests that this may explain why, in their own writing, Japanese learners of English are more comfortable using the present perfect with temporal adverbials than without. Their over-representation in EFL textbook exercises may mean that Japanese learners of English have internalised their use as lexical markers that trigger the use of the present perfect.

2.2.2.3 The progressive

In an in-depth, corpus-driven analysis of the progressive aspect comparable to Schlüter's (2002) analysis of the present perfect (see 2.2.2.2), Römer (2005) examined how the progressive is represented in the dialogues of two popular textbook series also designed for German secondary schools. Her study broadly follows Mindt's methodology described in 2.1.2.3. A noteworthy difference, however, is that Römer (2005) only examines occurrences of the progressives in the textbook passages intended to reflect spoken language use (printed dialogues, speech bubbles, transcripts of audio materials, etc.). By comparing these with how the progressive is used in everyday conversation among L1 speakers, her study is one of the few investigations of Textbook English to date that genuinely accounts for the fact that mode and register are likely to impact how such a grammatical construction is used in context.

For each occurrence of the progressive in her corpora, Römer (2005) surveyed a wide range of contextual features including tense forms, contraction, polarity, clause type, adverbial specification, verb lemma, subject and object of progressive verb phrases, as well as functional features including time reference, continuousness, repeatedness and framing (see 4.1.2 for details). Among other findings, Römer (2005: 244–245) reports that contracted forms of the auxiliary BE are under-represented among the progressive forms encountered in the textbook dialogues. Furthermore, she reports an

overuse of *he* and *she* in subject positions of progressives, whilst *I, it, we* and *they* are underused (Römer 2005: 246–248). With respect to the core functions of the progressive, Römer (2005: 260–266; see also Römer 2010: 22–24) notes that proportionally too few occurrences of textbook progressives convey the sense of "repeatedness". Römer (2005) also attempts to compare the results of some of her analyses across the most frequent verb lemmas; however, the textbook corpus surveyed being relatively small (108,000 words), such comparisons of the contextual use of specific verb lemmas in the progressive are inescapably explorative in nature.

2.2.2.4 Modals

Römer (2004) also applies a comparative corpus-based methodology following Mindt (see 2.1.2.3) to compare the frequencies, co-occurrence patterns and functions of modal verbs in authentic spoken British English and in the German secondary school EFL textbook series: *Learning English Green Line New*. As electronic versions of the coursebooks were unavailable, Römer (2004) decided against the compilation of a "pedagogic corpus" containing all the texts featured in the coursebooks, opting, instead, for the non-random selection of 32 texts from the textbook units in which one or more modals are specifically taught. Combined with all the grammar sections from the same textbook series and the content of a grammar book for the same level (*Learning English Grundgrammatik)*, the 32 texts were considered to represent "a sample of EFL textbook language – the kind of language pupils are exposed to in the EFL classroom" (Römer 2004: 190). Striking discrepancies are observed between the textbook data and the reference L1 corpus. For instance, whereas Römer's analysis of the spoken BNC1994 reveals that the modal *would* (and its contracted form *'d*) is the second most frequent modal, it only comes in fifth position in the textbook data (Römer 2004: 193). On the whole, in Textbook English, modals more frequently refer to ability than in naturally occurring conversation. Thus, for *could* and *may*, the meaning of possibility tends to be under-represented in the textbook data. Furthermore, *must* expresses an inference/deduction in over a third of the BNC concordance lines examined, yet this meaning is only very rarely featured in textbooks (Römer 2004: 194). Textbooks also tend to over-represent certain negated modals whilst others are never presented in a negated form in the textbook materials examined (Römer 2004: 194). Römer (2004) notes further mismatches between the two corpora in the context of questions and *if*-sentences featuring modals (Römer 2004: 195).

2.2.2.5 Conditionals

Frazier (2003) surveyed eight ESL textbooks for their coverage of hypothetical and counterfactual conditionals. The study concludes that the textbooks largely neglect hypothetical and counterfactual *would*-clauses that are removed from their presumed *if*-clauses and rarely present such clauses in larger units of discourse. To demonstrate the prevalence of such *would*-clauses, Frazier (2003) also conducted a quantitative

and qualitative analysis of 467 instances of *would*-conditionals in one written and two spoken corpora of American English. Frazier (2003: 451) observes that "much more often than not, conditional and hypothetical clauses with the modal *would* are not accompanied by *if*-clauses anywhere near them, much less in the same sentence". The author draws two major conclusions from his study. On the one hand, he encourages EFL/ESL textbook authors to "move away from the common practice of teaching hypothetical and counterfactual results in *would*-clauses only at the sentence level and only as adjacent to overt *if*-clauses" and on the other, he calls for a re-examination of the grammatical terminology used to describe *would*-clauses, starting with the term *conditional* itself (Frazier 2003: 464).

Taking Frazier's (2003) corpus-driven data as her basis, Yoo (2013) investigated how *would*-clauses are presented in five Korean high school EFL textbooks which occupy around 70% of the EFL textbook market share in Korea. She extracted all 253 occurrences of *would*-clauses from the textbooks and annotated these for various linguistic variables. Yoo also compared the lexico-grammatical patterns in which the *would*-clauses are embedded and compared their frequencies with Frazier's (2003) analysis of naturally occurring English corpus data. Yoo (2013) observes great disparities between the textbook sentences and naturally occurring language. For instance, she notes that despite its wide usage, all the textbooks explored fail to include the combination of *would* with the copula verb *seem* (Yoo 2013: 54), thus neglecting to expose learners to double hedging construction with a potentially highly useful pragmatic use. Moreover, no single example of an infinitive-*would* pattern is found in the textbook corpus (Yoo 2013: 81) despite attested frequent occurrence in naturally occurring English as demonstrated by Duffley's (2006) corpus study.

Gabrielatos' (2003; 2006; 2013) approach to evaluating the authenticity of *if*-sentences as described in EFL textbooks has already been presented in 2.1.2.4. These studies clearly demonstrate that a large proportion of *if*-sentences found in natural L1 speech and writing cannot be accounted for by the consensual typologies of conditionals typically taught in EFL/ESL textbooks. Remarkably, this is true even when taking account of all the rules and examples featured in the most advanced textbooks examined. Furthermore, Gabrielatos (2013: 158) notes that "ELT coursebooks generally adopt a naive and restricted approach to modal marking, as they tend to focus on central modals". On the basis of these pedagogically-driven corpus-based studies, Gabrielatos (2013: 155) draws the worrying conclusion that "the pedagogical information in the coursebooks, taken collectively, presented learners not only with a partial picture of the variety of types of conditionals and their respective morphosemantic features, but also a distorted one".

Using the same textbook corpus as for Römer's (2005) analysis of the progressive, Römer (2004a) also conducted a study on conditionals in Textbook English. In this

study, the focus lies on the sequences of clauses and tenses in conditional sentences and collocational patterns within *if*-clauses. Römer (2004a: 158) reports that a higher proportion of *if*-sentences begin with the *if*-clause in the authentic data than in the textbook corpus. The results also show that three tense form sequences are vastly over-represented in the textbooks as compared to the corpora of naturally occurring speech. These tense sequences correspond to what EFL textbooks and grammar books usually refer to as Type 1, Type 2 and Type 3 conditionals. Conversely, Römer (2004a: 159–160) demonstrates that the most frequent tense combinations in *if*-sentences in the spoken component of the BNC2014 (simple present + simple present), as well as a number of other frequent tense sequences appear to be significantly under-represented in Textbook English (for similar results, see also Gabrielatos 2003; 2006; 2013; Möller 2020; Winter & Le Foll forthcoming).

2.2.2.6 Reported speech

Applying a manual page-by-page approach (see 2.1.1.2), Barbieri & Eckhardt (2007) surveyed how reported speech is taught in seven popular ESL/EFL grammar textbooks. They report that the textbooks largely focus on indirect reported speech and find a general consensus on the 'backshifting rule' for pronouns, adverbials and tense. Whilst all the textbook authors seem to agree that, in general, the verb in the embedded clause should be "backshifted" to the past, there is no agreement as to which specific cases constitute exceptions to this rule. Barbieri & Eckhardt (2007) compared these results to two corpus-based studies on direct (Barbieri 2005) and indirect (Eckhardt 2001) reported speech in real language use, mobilising conversation and newspaper corpora as their data basis. They drew on the most striking discrepancies between the textbooks' "grammar rules" and the patterns of use that emerge from the authentic data to make ten suggestions to improve the authenticity of EFL textbook's portrayal of reported speech. Thought-provokingly, Barbieri & Eckhardt (2007) convincingly argue that indirect reported speech should not be taught as a transformation of direct speech (i.e., following the well-known backshifting rule), since the two constructions follow distinctive lexico-grammatical patterning and discourse functions and are used in different communicative situations and registers. Consequently, the authors suggest that indirect reported speech should be taught in the context of newspaper writing, whilst direct reported speech ought to be taught in the context of conversation. Barbieri & Eckhardt (2007) also make recommendations concerning the range of reporting verbs that ought to be associated with certain types of reported speech constructions in EFL/ESL textbooks and encourage textbook authors to highlight the grammatical patterns and discourse functions associated with less frequent tense sequences. Furthermore, they advocate the inclusion of informal quotatives such as *be like*, *go* and *be all* in textbooks, together with context regarding their discourse-pragmatic function and sociolinguistic associations.

2.2.3 Pragmatics

This section examines aspects of discourse and pragmatics in Textbook English. The selected studies are discussed in chronological order.

In English, doubt and certainty can be expressed in a variety of ways including using certain modal verbs, adjectives, tag questions, and specific intonation patterns, as well as paralinguistic and non-linguistic devices. The fact that many lexical markers of doubt and certainty are highly polysemous further adds to the complexity of the task. This motivated Holmes (1988) to conduct a survey on the coverage of lexical items commonly employed to express doubt in two EFL/ESL reference grammars and two coursebooks. To this end, Holmes (1988) compared the epistemic lexical items illustrated in the four textbooks with information on the range and frequency of the same lexical devices as retrieved from four different native corpora. In addition to these corpus-based comparisons, Holmes (1988) also referred to previous research findings and to native-speaker acceptability judgments to evaluate the choices of the pedagogical material designers. Holmes' (1988) study concludes that some textbooks paint an entirely misleading picture of epistemic modality compared to real-life English usage, whilst other textbooks neglect the topic altogether. Although corpus data clearly shows that when expressing doubt, native speakers of English do not confine themselves to modal verbs, Holmes (1988: 40) observes that the majority of textbook authors "devote an unjustifiably large amount of attention to modal verbs, neglecting alternative linguistic strategies for expressing doubt and certainty". As a result, she argues that textbooks ought to "present learners with alternative syntactic and lexical devices selected from those occurring most frequently in relevant spoken and written texts" (Holmes 1988: 40).

In academic writing, doubt is frequently expressed in the form of hedging. Hyland (1994) explored the representations of hedging devices in 22 EAP/ESP textbooks designed to help L2 English users acquire Academic English writing skills. Following a page-by-page approach (see 2.1.1.2), Hyland (1994) first drew a list of markers of uncertainty and tentativeness that a number of previous studies have found to be salient in academic writing and proceeded to manually check the EAP/ESP textbooks for evidence of coverage of these hedging devices. The evaluation of the textbooks' coverage of these devices is based on both the number of exercises devoted to these devices and the quality of the information provided on them (Hyland 1994: 244). Hyland's (1994: 250) study concludes that, in general, "the presentation of hedges in published [EAP/ESP] materials is not encouraging, with information scattered, explanations inadequate, practice material limited". Echoing Holmes' (1988) conclusion, Hyland (1994: 244) criticises the fact textbooks hardly present any alternatives to modal verbs for hedging.

For her cursory exploration of pragmatic information featured in English textbooks, Vellenga (2004) selected four EFL integrated skills textbooks and four ESL grammar textbooks that are frequently used in university settings in non-English-speaking countries. The study's methodology has already been detailed in 2.1.1.2. The results can only be considered exploratory but seem to point towards a general paucity of metapragmatic and metalinguistic information in EFL and ESL textbooks. The author notes that metalinguistic information is mostly presented in the form of imperative directives for learners to complete an activity in the textbook. Pronominal reference is often absent. Vellenga (2004) deplores the pragmatic inadequacy of the treatment of most speech acts in the textbooks explored. In particular, she points to the danger of providing unique speech act–grammatical form associations as they may prove misleading and restricting for learners. Based on informal observation, Vellenga (2004: n.p.) argues that the "distribution of speech act types across ESL and EFL textbooks did not appear to be patterned, nor based on frequency of speech act occurrence in natural language". Finally, interviews conducted with four experienced EFL/ESL teachers lead the author to the disconcerting conclusion "that textbooks do provide the majority of input, and that even professional teachers rarely have the time, inclination, or training to include supplementary pragmatic information in their lessons" (Vellenga 2004: n. p.).

Cheng (2007) examined a corpus of spoken English produced by competent English speakers from Hong Kong (the *Hong Kong Corpus of Spoken English*; Cheng, Greaves & Warren 2005) for instances of interruption in the form of initiation of simultaneous talk in conversations. Both the functions of the interruptions and their linguistic realisations were compared to phrases that eleven popular English textbooks used in Hong Kong secondary schools suggest are appropriate for interrupting. Cheng's (2007) results make clear that, by and large, the phrases suggested in the textbooks do not accurately reflect real-language use: the majority of the phrases taught in the textbooks (e.g., *Excuse me, but…, Sorry to interrupt…, If I could just come in here…, I want to say something, please)* do not occur even once in the 800,000-word corpus of naturally occurring Hong Kong English corpus queried.

Drawing on the same reference corpus, Cheng & Warren (2007) evaluated textbook authors' perception and presentation of strategies for monitoring and checking understanding in 15 upper secondary ESL/EFL textbooks. Their manual analysis of the pedagogic texts shows that the textbook authors emphasise the role of the listener in checking understanding, often providing example phrases encouraging the listeners to "seek clarification", "ask for repetition", "say they don't understand", etc. Only four of the 15 textbooks examined also propose strategies for the speaker to "check others' understanding" and "clarify". Contrary to textbook authors' intuition, the authentic conversation data reveal that the primary responsibility in ensuring that understanding has taken place rests, in fact, with the speaker. Cheng & Warren (2007)

also report that some of the phrases suggested by textbook authors to explicitly ask the listeners whether they have understood (e.g., *Are you with me?* and *Do you understand me?*) do not appear a single time in the *Hong Kong Corpus of Spoken English*. Other phrases such as *Is that clear?* only feature in the corpus' academic subcorpus and are only uttered by teachers speaking to students. Whilst the simple backchannel *okay* is found to be the third most frequent form of checking understanding in the corpus of real-life English, it is widely under-represented in textbook examples. Thus, Cheng & Warren (2007: 202) conclude "that textbooks contain language forms that are rarely, if ever, used in the real world and are overly influenced by academic genres". Most interestingly from a pedagogical point of view, they also conclude that "the most common forms found in the HKCSP [reference] corpus are both simpler and less explicit than those included in the textbooks" (Cheng & Warren 2007: 202).

In a similar vein and, again, relying on the same corpora, Cheng & Warren (2005; 2006) reveal that Hong Kong textbook authors' intuitions on forms used to express disagreement (Cheng & Warren 2005) and to give opinions (Cheng & Warren 2006) do not match real language use as documented in the *Hong Kong Corpus of Spoken English*.

In sum, the Textbook English studies surveyed in this section convincingly demonstrate that textbook authors' intuitions and portrayal of the functions and linguistic realisations of a wide range of learner-relevant speech acts largely fail to correspond to evidence from corpora of naturally occurring English. It is worth noting that, unlike the vast majority of corpus-based textbook language studies examined so far, the aforementioned studies carried out in the Hong Kong context rely on a reference corpus that consists of culture-specific spoken interactions between Hong Kong Chinese speakers and speakers of languages other than Cantonese in a range of communicative situations that English learners are likely to be confronted with. Thus, contrary to many of the previously described Textbook English studies that rely on general L1 corpora of sometimes doubtful relevance to anyone but highly advanced learners of English (see Tono 2004), the *Hong Kong Corpus of Spoken English* can reasonably be considered to constitute secondary school target learner language. As a result of these apparent mismatches between textbook and target learner language, Cheng & Warren (2007: 200) warn of "an urgent need for some realignment in learning and teaching materials in terms of language forms and functions selected and presented to language learners".

2.2.4 Spoken grammar

The final section of this literature review focuses on how spoken grammar is represented in EFL/ESL textbooks. Attitudinal research indicates that learners of English are generally very interested in possessing at least receptive knowledge of

spoken grammar, whereas teachers and textbook authors are more divided on the subject (Timmis 2003). Although he refrains from citing specific textbooks, Timmis (2003) highlights a number of spoken grammar features that he considers to be largely ignored by ELF materials designers. These include the *get*-passive, discourse dimensions of past-aspect choices and certain reported speech forms, in addition to other grammatical features which analyses of L1 speech corpora have shown to be very frequent, such as heads, tails and ellipsis structures (see, e.g., Aijmer 2002; Carter & McCarthy 2006b; Carter & McCarthy 1995; McCarthy & Carter 1995; Carter, Hughes & McCarthy 1998; McCarthy 1998; Biber & Quirk 2012; Biber et al. 2021). This claim would certainly merit quantitative verification and, in fact, a number of Textbook English studies are concerned with the specific linguistic features of spoken English.

One original methodological approach to evaluating the linguistic features of spoken interactions in EFL textbooks was already presented in 2.1.2.5. As a reminder, Gilmore's (2004) evaluation of the authenticity of spoken textbook language involved the comparison of re-creations of communicative situations simulated in textbooks with the textbook dialogues themselves. Nine discourse features were selected for comparison: lexical density, false starts, repetition, pauses, terminal overlap, latching, hesitation devices and backchannels. Gilmore (2004) reports that the 'authentic' enactments of the dialogues (see caveats in 2.1.2.5) are almost twice as long as their textbook counterparts. Additionally, the lexical density of textbook dialogues is higher than that of authentic conversations (though it may be argued that the difference, based on 200-word samples from only seven examples, does not appear to be great, nor is this claim backed by the results of any statistical significance testing). In particular older textbooks were found to hardly feature any of the discourse features typical of spontaneous spoken interactions.

In a page-by-page survey, Cullen & Kuo (2007) focused on explicit mentions of features of spoken grammar in 24 global EFL textbooks covering all levels from beginner to advanced. Drawing on specific examples from Carter (2004), they investigated three categories of features of spoken grammar. Cullen & Kuo (2007: 361) conclude that "where spoken grammar is dealt with at all, there tends to be an emphasis on lexico-grammatical features, and common syntactic structures peculiar to conversation are either ignored or confined to advanced levels as interesting extras". One could argue, however, that such structures are more likely to be taught implicitly than in an overt manner and that, as a consequence, a corpus-based textbook study that includes the written dialogues and transcripts of the listening materials accompanying the textbooks may paint a rather different picture.

The results of the studies outlined above confirm that, despite a strong focus in syllabi on speaking skills and communicative language learning, textbooks seemingly

continue to present a misleading picture of spoken language. Barbieri & Eckhardt (2007: 321) reflect that even more recent publications "neglect important and frequent features of the language spoken by real language users, present a patchy, confusing, and often inadequate treatment of common features of the grammar of the spoken language, and, in sum, do not reflect actual use". However, this section has shown that studies on representations of spoken English in EFL/ESL textbooks have either only focused on individual linguistic features (e.g., Römer on *if*-conditionals [2004; see 2.2.2.5] and the progressive [2005; see 2.2.2.3]), on explicit mentions of spoken grammar features only (e.g., Cullen & Kuo 2004), or on a very small sample of textbook dialogues (e.g., Gilmore 2004).

2.3 Conclusion

This review leaves no doubt that, as suspected by many (former) learners of English as a foreign language in instructional contexts (see 1.1), Textbook English does indeed constitute a distinct variety of English that, in many respects, differs substantially from real-life, naturally occurring English. Section 2.2.2.5 highlighted the fact that some grammar rules promulgated in textbooks are, in fact, not even respected in the extended written passages of the textbooks themselves; thus, pointing to striking intra-textbook inconsistencies and, more generally, to a genuine gap between prescriptive grammars of English and real-life usage.

Section 2.2.3 showed that the results of most studies examining pragmatics in Textbook English stress that textbooks are not providing learners with the right kind of input to develop their pragmatic competence. In spite of some improvements found in studies comparing older with more recent publications (e.g., Gilmore 2004; Jiang 2006; Usó-Juan 2008), critics argue that newer publications do little more than simply lengthen the list of linguistic structures to be used in the context of specific speech acts, yet provide next to no contextual information as to their use in real language use (Usó-Juan & Martínez-Flor 2010: 426).

In terms of methods, we have seen that linguists have, thus far, mostly strived to compare Textbook English to naturally occurring English as produced by native speakers; however, even comparisons of local ESL textbooks with that of English produced by local proficient ESL speakers have shown that the language input learners obtain from their textbooks remains far removed from what they are expected to later engage in outside the classroom (Cheng 2007; Cheng & Warren 2005; 2006; 2007).

In order to better grasp this apparent mismatch between the language of pedagogical materials and target learner language, it is important to consider the factors that may contribute to this gap. The first argument that textbook authors would presumably advance is that their task is to simplify real-life language use to make it accessible to

language learners. In fact, a distinct limitation of many of the studies surveyed as part of this literature review is that they examine Textbook English as one single variety of English, ignoring potential variation related to the different proficiency levels of the textbooks. Whilst there is no doubt that proficiency level must be accounted for in future descriptions of Textbook English (see 3.1), advocates of usage-based L2 instruction models would counter that the most important factors in the construction of Textbook English ought to be the relative frequencies of occurrence and salience of linguistic features (see 1.5). In reality, however, it would appear that the majority of textbook authors still rely on long-established grammar conventions and/or their intuition, rather than on (corpus-based) insights from actual language usage (see 1.8). Many of the conclusions of the Textbook English studies summarised in this chapter thus reminded us that: "Tradition, even if it is most venerable, cannot serve as a substitute for research" (Mindt 1997a: 41). This was particularly obvious in the studies that compared the coverage of key lexical items in different textbooks targeted at the same proficiency level and found very little agreement across different textbook series (e.g., Koprowski 2005; see 2.2.1.2).

Furthermore, many of the studies surveyed as part of this literature review have concluded that textbooks tend to present lexico-grammatical patterns as if they were generalisable across all registers, thus failing to acknowledge crucial differences between different production modes, text types and discourse-context-specific uses (see also Barbieri & Eckhardt 2007: 321). The studies summarised in 2.2.3 and 2.2.4, in particular, suggest that the language of EFL textbooks is predominantly based on norms pertaining to written text registers. Indeed, to a large extent, textbook dialogues fail to account for the processing conditions of spontaneous spoken interaction (e.g., lack of planning, reciprocity, shared environment). The resulting disregard of syntactic and lexico-grammatical features that are typical of unplanned speech therefore also contributes to this prevailing lack of fit between many aspects of Textbook English and real-life language use. At the same time, however, this criticism can be turned on its head. As this chapter has shown, numerous Textbook English studies suffer from severe methodological limitations. Crucially, many of the conclusions drawn on the basis of comparisons between Textbook English and naturally occurring English are likely flawed due to the use of inappropriate or only partially suitable reference corpora (e.g., comparing textbook dialogues to the entire BNC1994).

In spite of their limitations, however, the studies outlined in this chapter, collectively, provide substantial evidence for the frequently idiosyncratic use of specific lexico-grammatical features in Textbook English. They reveal pedagogically questionable gaps between the language input textbooks provide and what learners are expected to eventually engage with. However, in spite of the considerable scope of linguistic features explored thus far, our understanding of Textbook English remains patchy –

not least because no study has yet attempted to provide a comprehensive linguistic description of Textbook English across a broad range of linguistic features and registers, based on appropriate reference corpora and a sufficiently large corpus of EFL textbooks. The present study sets out to contribute to bridging this gap by investigating a broad range of lexico-grammatical features in a large corpus of contemporary EFL textbooks widely used in secondary schools in France, Germany and Spain. The following chapter explains how this literature review informed both the methodological approach of the present study (3.1–3.2) and the design of the corpora used in its comparative corpus-based analyses (3.3).

3 Research aims and data

One does not study all of botany by making artificial flowers.
❖ John McH. Sinclair (1991: 6)

The present chapter begins by drawing conclusions from the studies discussed as part of the literature review in Chapter 2. On the basis of these insights, the overarching methodological framework for the present study is briefly outlined in 3.2. Four broad sets of research questions are formulated in 3.2.1. These concern the linguistic nature of Textbook English and the various factors that may mediate variation within this variety of English. They will be further specified in each of the four analysis chapters that follow. Since these analysis chapters, Chapters 4–7, can be read as individual studies, they each begin with a set of specific research questions and their own methods sections. However, the analyses presented in Chapters 4–7 are united by both the overarching research aims and questions spelt out in 3.2 and by the fact that they rely on the same corpus data. The second half of the present chapter outlines the various decision-making processes involved in the compilation of these corpora and provides key statistics on their compositions.

3.1 Insights from the literature review

Taken together, the studies surveyed in Chapter 2 provide valuable insights into "the kind of synthetic English" (Römer 2004b: 185) that EFL learners are exposed to via their English textbooks. However, the literature review revealed some problematic aspects that have commonly been neglected in past endeavours to study the language of textbooks and which the present thesis aims to address.

First, throughout the literature review, concerns were raised as to the suitability of the reference corpora used in comparative corpus-based analyses. Whilst some have argued that English native speaker standards are not suitable for most EFL learners (for more on this, see 3.3.2.1), the more pressing issue resides in the fact that many of the reference corpora used in previous Textbook English studies do not match the communicative aims and/or target audiences of the textbook texts. Many of the surveyed studies relied on general English corpora such as the British National Corpus (BNC), which is made up of 90% written language, mostly penned by professional writers with an adult readership in mind and edited by professionals for publication. Thus, the present study aims to find and, if necessary, compile the most appropriate reference corpora possible for maximally meaningful comparisons.

Second, though it has long been established that situational characteristics of texts are a major driver of linguistic variation (see, e.g., Biber 2012; Gray & Egbert 2019; Goulart et al. 2020), Chapter 2 repeatedly showed that potential register differences between the various types of texts typically featured in school EFL textbooks have largely been brushed aside. Given that school EFL textbooks may feature, for example, extracts of a short story, a dialogue, instructions, and exercises on any double page, the present thesis hypothesises that Textbook English cannot be meaningfully examined without taking a register-based approach. Up until now, however, register variation within EFL textbooks has largely been ignored (but see Miller 2011 for an exception with respect to university-level ESL textbooks). In the few cases where register has been considered in the analysis of EFL textbooks, the focus has almost exclusively been on representations of spoken language, e.g., Mindt (1987; 1995a) and Römer (2004a; 2005) who compared the dialogues of secondary school EFL textbooks to corpora of spoken and pseudo-spoken native speaker English. However, to the author's best knowledge, other sources of register variation in school EFL textbooks have yet to be explored. Other registers of interest in the analysis of secondary school EFL textbooks may include fiction, task instructions & explanations, informative texts, etc. (see 3.3.1.4).

Another frequently neglected aspect concerns interactions between the frequencies of individual linguistic features. This is important because usage-based approaches to language acquisition (see 1.4–1.5) postulate that the co-occurrence information that learners perceive in language input:

> is stored as points in a multi-dimensional space at coordinates, and that speakers process this stored linguistic information in ways that allow them to identify (under certain conditions and defined by various types of frequency occurrences) abstract linguistic patterns" (Rautionaho & Deshors 2018: 229).

Thus, whilst some influential studies have helped us to understand how EFL/ESL learners can be misled by their textbooks to make unidiomatic use of specific linguistic features (e.g., the progressive aspect; Römer 2005), only a multivariable approach can paint the full picture as to how Textbook English – as a whole – differs from the English that English learners will later encounter and be expected to produce outside the classroom. In fact, even within a case-study approach describing and/or evaluating the representation of a single linguistic feature in Textbook English, potential interactions between different variables ought to be taken into consideration. For instance, if a study reports that certain lexical verbs in the present progressive are under-represented in textbook dialogues as compared to naturally occurring conversation among L1 speakers, this could mean that these verbs are, overall, under-represented in EFL textbooks, across all tenses and aspects. It could, however, equally mean that the present progressive is under-represented as compared to other tenses and aspects. Alternatively, it could point to a genuine under-representation of a specific combination of tense, aspect and verb type. Similarly, if

such a study additionally reports that progressives with contracted auxiliary forms are under-represented, the question arises as to whether, in natural speech, present progressives with these particular verbs are proportionally more likely to feature in contracted forms than other verbs in the progressive, in which case these two findings can be accounted for by one and the same phenomenon. The problem is that Textbook English studies, to date, have not accounted for such possible interactions between the linguistic variables that they have measured, compared and reported.

As already hinted at by these examples, such multi-variable analyses of corpus data call for appropriate statistical methods. In the methodological review in Chapter 2, however, it was reported that statistical tests have only rarely been conducted in the context of quantitative comparative corpus-based analyses of Textbook English. When they have, chi-squared tests have been favoured (see, e.g., Römer 2005: 60). These statistical tests, together with other tests popular in corpus linguistics such as log-likelihood and Fisher's exact tests (Brezina 2018: 112–115), can help researchers conclude with a certain degree of certainty that frequencies observed in textbook language differ from the probabilistically expected frequencies (drawn from, e.g., a reference corpus) because of genuine differences between the corpora, rather than due to random, chance variation (Levshina 2015: 201–213; Wallis 2020: chap. 8). Many studies that report tests such as chi-squared tests, however, only report whether the test returned a p-value below a pre-defined threshold (usually 0.05, corresponding to 5% probability that the same test result or a more extreme one could have been obtained if there were no actual difference between the compared populations). Here, the problem is twofold. First, p-values do not inform the reader as to how large (let alone: relevant!) these supposed differences actually are and, second, they are dependent on sample size. Thus, given very large corpus datasets, statistically significant results (i.e., with very small p-values) will almost inevitably be returned (Baroni & Evert 2009: 787; Gries 2005; or simulate some data to observe this effect on https://shinyapps.org/apps/p-hacker/ [Schönbrodt 2016]). With small datasets, by contrast, only extremely large differences between two sets of frequencies will return significant results. In other words, studies with small sample sizes are often underpowered and therefore cannot be used to reliably detect anything but huge effects (Winter 2019: 171–175). Another issue is that when conducting individual tests on potential differences in the frequencies of many different linguistic variables, p-values ought to be corrected to account for multiple comparisons (Wallis 2020: 274–275; Winter 2019: 175–177). Again, so far, this has rarely been done in corpus-based analyses of Textbook English.

On a related matter, previous quantitative corpus-based studies of textbook language have usually been undertaken at the corpus level (rather than at the textbook volume, chapter, unit or individual text level), thus implicitly assuming that Textbook English is a homogenous variety of English in which the linguistic features under study are

dispersed evenly within the textbook corpora. Just like learner Englishes have been shown to vary across different registers, tasks, proficiency levels, individual learners, etc., so can Textbook English be expected to vary across different textbook series, targeted textbook audience/instructional setting, proficiency levels, text registers, etc.

3.2 The present thesis

The present study is based on Mindt's comparative, corpus-based methodology (see 2.1.2.3). As explained in the previous chapter, such an approach requires the compilation of two corpora: a textbook corpus and a reference corpus of naturally occurring English representative of the target language students can be expected to aspire to. In addition, various methods that have been proposed to account for the multifactorial/-variate and multilevel structure of learner corpus data (Gries 2013; 2018; Gries & Deshors 2020; Paquot & Plonsky 2017; Möller 2017; Wulff & Gries 2021) have inspired the methodological approaches followed in the present study in an attempt to account for the nested nature of textbook language data. Chapters 6 and 7, in particular, attempt to model the potential impact of the different text registers typically found in EFL textbooks, their varying target proficiency levels, and any potential idiosyncrasies of textbook authors, editors or publishers, as well as how these variables may interact with each other.

3.2.1 Research aims and questions

In sum, the present study attempts to describe both the linguistic specificities of Textbook English as a variety of English, as well as to model its internal variation. More specifically, the project addresses the following broad research questions:

1. How homogenous is Textbook English as a variety of English? Which factors mediate intra-textbook linguistic variation?
2. To what extent are French, German and Spanish secondary school pupils confronted with varying English input via their textbooks?
3. To what extent is the language of current EFL textbooks used in secondary schools in France, Germany and Spain representative of 'real-life' English as used by native/proficient English speakers in similar communicative situations? To what extent are some registers more faithfully represented than others?
4. Which (clusters of) lexico-grammatical features are characteristic of Textbook English? To what extent are these stable across entire textbook series? To what extent are some of these defining features specific to certain proficiency levels?

This study thus aims to provide an empirical, multi-variable description of the language of a large sample of secondary school EFL textbooks. As such, it is hoped that it can contribute to raising awareness of what constitutes the main variety of English that secondary school students are formally exposed to: "Textbook English". Ultimately, both the results and some of the methods employed may be used to evaluate future EFL teaching materials. In addition, the results may help EFL

teachers, textbook authors and editors to improve existing teaching materials in order to better equip learners with the necessary linguistic skills to succeed in the English-speaking world outside the classroom.

3.2.2 Open Science statement

Another important insight from the methodological part of the literature review (Chapter 2) is that, to the author's best knowledge, no Textbook English study published so far has included (as an appendix or supplementary materials) the data and code necessary to replicate the published results. This means that it is very difficult to evaluate the reliability or robustness of the results reported. Granted, a major issue in (corpus) linguistic research is that it is often not possible for copyright or, when participants are involved, data protection reasons to make linguistic data available to the wider public. However, both research practice and the impact of our research can already be greatly improved if we publish our code or, when using GUI software, methods sections detailed enough to be able to successfully replicate the full procedures. This step can enable others to conduct detailed reviews of our methodologies and conceptual replications of our results on different data.

Aside from data protection and copyright regulations, there are, of course, many reasons why researchers may be reluctant to share their data and code (Berez-Kroeker et al. 2018; McManus 2021). It is not within the scope of this thesis to discuss these; however, it is clear that, in many ways, such transparency makes us vulnerable. At the end of the day: to err is human. Yet, the risks involved in committing to Open Science practices is particularly tangible for researchers working on individual project, like the present author on this doctoral thesis, who have had no formal training in data management or programming and have therefore had to learn "on the job". Nonetheless, the author is convinced that the advantages outweigh the risks. Striving for transparency helps both the researchers themselves and others reviewing the work to spot and address problems. As a result, the research community can build on both the mishaps and successes of previous research, thus improving the efficiency of research processes and ultimately contributing to advancing scientific progress.

It is with this in mind that the author has decided, whenever possible, to publish all the raw data and code necessary to reproduce the results reported in the present thesis following the FAIR principles (i.e., ensuring that research data are Findable, Accessible, Interoperable and Reusable; Wilkinson et al. 2016). For copyright reasons, the corpora themselves and annotated corpus data in the form of concordance lines cannot be made available. However, the outcome of both manual and automatic annotation processes is published in tabular formats in the Online Appendix. These tables allow for the reproduction of all the analyses reported on in the following chapters using the reproducible data analysis scripts also published in the Online Appendix and on GitHub (https://github.com/elenlefoll/TextbookEnglish). In all

chapters of this thesis, full transparency is strived for by reporting on how each sample size was determined and on which grounds data points were excluded, manipulated and/or transformed. Most of these operations were conducted in the open-source programming language and environment R (R Core Team 2020). Most of the data processing and analysis scripts therefore consist of R notebooks. These were rendered to HTML pages (viewable in the Online Appendix) thus allowing researchers to review the procedures followed without necessarily installing all the required packages and running the code themselves. These scripts also feature additional analyses, tables and plots that were made as part of this study but which, for reasons of space, were not reported on in detail in the present thesis. Whenever additional software or open-source code from other researchers were used, links to these are also provided in the Online Appendix (in addition to the bibliographic references in the corresponding sections of the thesis).

3.3 Corpus data

Having outlined the overarching aims and research questions of the present thesis (3.2.1) and its commitment to Open Science (3.2.2), the second half of this chapter describes the corpora drawn upon to explore the characteristics of Textbook English as a variety of English.

The first part (3.3.1) explains the rationale behind the many decision processes involved in the creation of the Textbook English Corpus (TEC), e.g., the selection of the textbooks to be included in the corpus (3.3.1.1) and more technical aspects of the corpus compilation process such as the software used to process scans of physical copies of the textbooks obtained in print and those used to process digital copies of various formats (3.3.1.2), as well as the XML syntax used for the mark-up of textbook metadata (3.3.1.3) and register annotation (3.3.1.4).

The second part (3.3.2) is devoted to the three reference corpora used to compare and contrast Textbook English with naturally occurring 'real-life' language deemed to be representative of the kind of English that secondary school L2 English learners can be expected to understand and produce once they have completed their secondary school education. The reasoning behind the choice of the Spoken BNC2014 and how the untagged XML version of the corpus was processed for the purposes of this study are explained in 3.3.2.1. This section also includes a brief excursus on the choice of British English as the reference norm for the present analyses. The remaining two reference corpora – the Youth Fiction Corpus and the Informative Texts for Teens Corpus (abbreviated to Info Teens) – were compiled especially for the present project. Their composition and respective corpus building processes are described in 3.3.2.2 and 3.3.2.3.

3.3.1 The Textbook English Corpus (TEC)

The present study aims to examine the English language content EFL learners are exposed to in secondary school settings. To conduct a corpus-based analysis of this input, it is necessary to compile a pedagogic corpus, which Hunston (2002: 16) defines as:

> A corpus consisting of all the language a learner has been exposed to. For most learners, their pedagogic corpus does not exist in physical form. If a teacher or researcher does decide to collect a pedagogic corpus, it can consist of all the course books, readers, etc., a learner has used, plus any tapes etc they have heard. […] It can also be compared with a corpus of naturally occurring English to check that the learner is being presented with language that is natural-sounding and useful.

Collecting data for a pedagogic corpus as defined by Hunston is undoubtably a highly ambitious project. Although not explicitly mentioned, her definition implies that it should also include all teacher-student and student-student interactions in the L2 and would thus be specific to each and every class group and, even, learner. If, however, in "input-impoverished EFL context[s]" (Meunier & Gouverneur 2007: 122), textbooks do indeed account for such a large proportion of the learner language input, it follows that the textbooks themselves can be considered as a kind of "learner input corpus" (cf. Gabrielatos 1994: 13). In 1.6, textbooks were shown to be a key source of language input in school EFL classroom settings. This formed the starting point for investigating the language of EFL textbooks used at lower secondary school level in France, Germany and Spain. In designing and compiling any corpus, several aspects must be carefully considered. These include the corpus specification, the data collection sampling frame and considerations pertaining to corpus size, representativeness, and balance, all which are explained in the following sections.

3.3.1.1 Selection of textbooks

No matter how large, corpora tend to represent only a sample of a target population, with few exceptions such as corpora of individual authors' complete published works. In this study, however, the target population is defined as the English language content of all the textbooks from which all lower secondary school students in France, Germany and Spain were learning English as a second or foreign language between 2016 and 2018. Since the school systems are organised differently across the three countries of interest, lower secondary school is defined here for comparison as the equivalent to the German *Sekundarstufe I* or ISCED 2 (Unesco 2012; 2015), i.e., the stage where pupils are usually expected to be aged between 11–12 to 15–16 years. In most OECD countries, this period coincides with compulsory secondary education. The corresponding educational levels and year groups for France, Germany and Spain are displayed in Table 3 (data from Fournier, Gaudry-Lachet & DEPP-MIREI 2017). To compare textbooks aimed at similar levels and year groups across these different educational systems, an additional universal "country-neutral" textbook level variable is used throughout this study (see first column of Table 3). Textbooks for more

advanced secondary school English L2 learner were not included in the corpus because textbooks are used more sparingly beyond the first four to five years of secondary school EFL instruction. Although still present in many European classrooms, they are often supplemented with other "real-life" materials (see, e.g., Leroy 2012: 72 for the French context).

Table 3: The levels of the Textbook English Corpus (TEC)

TEC Level	**France**		**Germany**		**Spain**		**Pupil age (approx.)**
A	Collège	6e	Sekundarstufe I	5. Klasse	Educación Primaria	6°	11
B		5e		6. Klasse	Educación Secundaria Obligatoria (ESO)	1° ESO	12
C		4e		7. Klasse		2° ESO	13
D		3e		8. Klasse	ESO Secundo Ciclo	3° ESO	14
E	Lycée	2e		9./10. Klasse		4° ESO	15

Since the French, German and Spanish educational authorities do not prescribe the use of any particular textbooks in state schools, a vast number of different textbook series from a range of publishers are currently in use. For a textbook corpus to capture the full variability of lower secondary school EFL textbooks used in France, Germany and Spain, it would have to include all the textbooks in use, including possibly some older or little-known editions favoured by individual schools or teachers. However, the principle of representativeness, as described by Biber (1993a), also implies that the corpus ought to be representative of the textbook language to which as many pupils as possible are exposed, so that conclusions drawn from the sample of textbooks contained in the corpus may be confirmed in a larger sample of textbooks. This, in turn, implies that the most popular textbooks used in the majority of classrooms ought to be included in the corpus. Since textbook sale figures are not publicly available, informal surveys were conducted with local teachers (in EFL teacher Facebook groups), bookshop assistants and publisher representatives to establish a list of the most widely used school EFL textbooks in France, Germany and Spain.

Table 4 summarises the results of these informal market surveys, which revealed differences between countries in school textbook market dynamics. In Germany, the textbook market is dominated by three publishers (Klett, Cornelsen and Diesterweg), which each offer one major English textbook series per school form (usually: *Hauptschule, Realschule, Gesamtschule* and *Gymnasium*) and which, to a lesser extent, may be adapted to match the requirements of specific *Länder* (note that in Germany, education falls under the responsibility of the *Länder* as opposed to the federal ministry). By contrast, France has a centralised national educational system.

That said, schools and teachers are also free to choose whether to use textbooks at all, and if so, which. In general, there is a tendency to be more critical of textbooks, with some trainee teachers instructed not to use any commercial textbooks, or at least to design their own lesson units selecting suitable materials from a range of textbooks and authentic materials, rather than religiously following one series. In practice, however, it would appear that the majority of English teachers in lower French secondary schools do largely rely on one textbook per year group (Leroy 2012: 62) and, in fact, the school textbook market in France continues to show record growth in spite of the concerns voiced by critics (Syndicat national de l'édition 2021).

Whilst the textbooks used in French and German secondary schools are usually published in France and Germany, Spanish schools, teachers (and parents?) seem less convinced of the quality of their locally published textbooks and, as a result, the textbook market is largely dominated by Anglo-Saxon publishers. It may be speculated that such "imported textbooks" are favoured for the same reasons as they are in Southeast Asia, where Dat (2008) reports that they are perceived as being more visually and thematically appealing, linguistically accurate, and systematic in their pedagogical approaches compared to "domestic textbooks". Since the only Spanish publisher of school English textbooks featured in the informal survey list did not wish to contribute textbooks to the present project and was technically unable to sell digital textbooks without a Spanish ID number, only Anglo-Saxon published textbook series could be included in the Spanish textbook sub-corpus. These textbook series are generally the result of a core "global coursebook" (Gray 2002: 151, though the article dates from a time where this phenomenon was largely restricted to adult EFL/ESL textbooks series) sold to a number of target countries with "differentiated supplementary materials […] often written by local authors with specific local knowledge […] to give the teachers 'a better fit'" (Gray 2002: 165).

Table 4: Most widely used lower secondary school textbook series (and publisher in brackets) according to the informal market surveys conducted with teachers, bookshop assistants and publishers in France, Germany and Spain

France
Hi there! (Bordas)
Join the team (Nathan)
New Enjoy (Hatier)
E for English (Didier)
Piece of Cake (Le Livre Scolaire)
New Connect (Hachette)
Spain
High Achievers (Richmond)
Fast Track (Richmond)
Action! (Burlington)
Real English (Burlington)
English in Use (Burlington)

English in Mind for Spanish Speakers (Cambridge University Press) English File (Oxford University Press)
Germany
Gymnasium
Green Line (Klett) Access G (Cornelsen) Camden Town (Diesterweg)
Gesamtschule
Orange Line (Klett) Lighthouse (Cornelsen)
Hauptschule
Blue Line (Klett)
Realschule
Red Line (Klett)

In designing the sampling frame of the Textbook English Corpus (TEC), the aim was to select three recently published textbook series per country, ideally by three different major publishers, from the list compiled from the informal surveys (see Table 4). The selection was based on the following opportunistic criteria:

1. Availability of the textbooks in
    a. Text/PDF format
    b. Other digital formats and
    c. Print (in that order)
2. Price

Price was particularly relevant for French and Spanish textbooks as some publishers were only willing to sell digital textbooks in bundles of 20 textbooks or more. Availability may seem an odd criterion in an interconnected globalised world, but in some cases buying even a single digital copy of a Spanish textbook requires a valid Spanish ID number. Digital textbook formats also raised technical issues. Whilst PDF textbooks are relatively easy to convert to text using standard optical character recognition (OCR) software, many digital textbooks are only available as complex flash files designed for use with smartboards and/or tablets. These had to be converted to PDF on a page-by-page basis (though this was automated with a script) before they could be converted to text. Finally, two textbook series were obtained in print and scanned to PDF for further OCR processing.

The textbooks included in the TEC are listed in Table 5 (the full bibliographic metadata is available on doi.org/10.5281/zenodo.4922819). To ease comparisons across different educational systems, whenever possible, five textbooks per series were included. As a result, two French textbook series designed for use in *collèges* (corresponding to the first four years of secondary school education, see Table 3 – note, also, that French school years are counted backwards) are complemented with

a fifth textbook aimed at first year *Lycée* students in *Seconde* (2e; see Table 3) from the same publisher. This was not possible in the case of the most recent textbook series, *Piece of Cake*, since this relatively new publisher had, as of September 2018, not yet entered into the *lycée* textbook market. In terms of its marketing concept, *Piece of Cake* is rather different from the other textbook series featured in the TEC since it was co-authored by, at the time of writing, over 100 school English teachers, is published under a Creative Commons license for free modification and use, and is available online in its entirety for free on https://www.lelivrescolaire.fr. Though it was still relatively new when the TEC was compiled, the informal market study revealed it to be a very popular series in French secondary schools already.

Every effort was made to also include all the (tran)scripts of the audio and video materials belonging to the textbooks of the corpus. When they were not provided by the publishers themselves, this involved trawling through teachers' books and textbook home pages to access the materials. Unfortunately, transcripts could not be sourced for the older version of *Green Line* (Klett, 2006–2009 edition) or *Achievers* (Richmond). None of the textbooks' accompanying workbooks were included. This decision was based on pragmatic time and resource constraints and is justified by the fact that many schools do not require parents to buy the workbooks and many teachers do not, or only rarely, use them.

Though the TEC was constructed as a balanced corpus with three textbook series for each country of use, as shown in Fig. 1, the total word count is not spread equally across the three "national" subcorpora. In terms of quantity, the French textbooks feature considerably less English input than the German and Spanish ones.

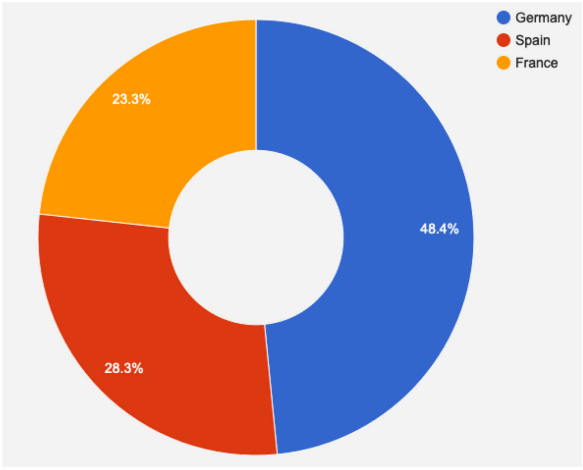


Fig. 1: Proportion of tokens in the three "national" subcorpora the TEC (as displayed by Sketch Engine; Kilgarriff et al. 2014)

Of the textbooks featured in the TEC, the only series advertised as corpus-informed is *English in Mind* from Cambridge University Press. However, many of the other textbook series claim to include a large proportion of "authentic materials".

Table 5: Composition of the Textbook English Corpus (TEC) (the full bibliographic metadata is available on doi.org/10.5281/zenodo.4922819)

Country of use	Publisher	Textbook series	Volume	Level	Publication date
France	Bordas	*Hi There*	6ème	A	2012
			5ème	B	2013
			4ème	C	2014
			3ème	D	2015
		New Mission	2nde	E	2014
	Nathan	*Join the Team*	6ème	A	2010
			5ème	B	2011
			4ème	C	2012
			3ème	D	2013
		New Bridges	2nde	E	2010
	Le Livre Scolaire	*Piece of Cake*	6ème	A	2017
			5ème	B	
			4ème	C	
			3ème	D	
Germany	Klett	*Green Line*	1	A	2006
			2	B	
			3	C	2007
			4	D	2008
			5	E	2009
	Klett	*New Green Line*	1	A	2014
			2	B	2015
			3	C	2016
			4	D	2017
			5	E	2018
	Cornelsen	*Access G*	1	A	2013
			2	B	2014
			3	C	2015
			4	D	2016
			5	E	2017
Spain	Richmond	*Achievers*	A1+	A	2015
			A2	B	
			B1	C	
			B1+	D	
			B2	E	
	Cambridge University Press	*English in Mind*	Starter	A	2010
			1	B	
			2	C	
			3	D	2011
			4	E	
	Oxford University Press	*Solutions*	Elementary	A	2014
			Pre-Intermediate	B	2016
			Intermediate	C	2017
			Intermediate Plus	D	2017

3.3.1.2 Textbook processing

The nine selected textbook series, or 42 textbook volumes (see Table 5), were processed so as to include as much of the textual content as possible. All PDF files were processed with high-performing OCR software (ABBYY FineReader 14 Corporate). The results were saved as text (.txt) files for future processing with corpus analysis software. All non-text elements such as images, symbols, font specifications, etc. were discarded.

3.3.1.3 Corpus mark-up: headers

To ensure maximum compatibility across operating systems and software, the corpus files were saved with Unicode UTF-8 encoding. In keeping with standard corpus practice, eXtensible Markup Language (XML) was used for the markup and annotation. However, in line with Atkins et al.'s (1992) advice to aim for "a level of mark-up which maximizes the utility value of the text without incurring unacceptable penalties in the cost and time required to capture the data" (Atkins, Clear & Ostler 1992: 9; see also Hardie 2014), the standards usually advocated for XML corpus mark-up, such as the Corpus Encoding Standard (CES; Ide 1996) and the Text Encoding Initiative (TEI; Burnard, Lou & Bauman 2021), were deemed unnecessarily detailed and too labour intensive. Since it was not relevant to the research question at hand, the textbooks' structural and formatting elements (e.g., paragraphs, font formats, section dividers, etc.) were not annotated and the design of the XML schema was therefore kept as simple as possible.

The metadata associated with each textbook was encoded in a simple XML header at the start of each textbook file, e.g.:

> <doc sign="POC4" series="Piece of cake" level="C" publisher="Livre scolaire" year="2017" country="France">

Each file header includes:

i. A unique file name (doc sign)
ii. The name of the textbook series
iii. The textbook proficiency level (according to the country-neutral scale introduced in 3.3.1.1 and Table 3)
iv. The publisher
v. The date of publication
vi. The country in which the textbook is used (i.e., France, Germany or Spain).

This simple metadata markup schema makes it possible to restrict corpus searches to subsections of the corpus, choosing for instance one or more level(s), publisher(s), series, country, a publication date range, or any combination of these parameters in off-the-shelf corpus software such as Sketch Engine (Kilgarriff et al. 2014), as well as via custom scripts in, e.g., R or Python.

3.3.1.4 Register annotation

A major pedagogical implication to emerge from corpus-based research is "the centrality of *register* for studies of language use" (Conrad & Biber 2013: 334, emphasis added). Thus, before outlining the register annotation process, a brief excursus on register is called for. Indeed, corpus studies have consistently shown that, at all linguistic levels, lexico-grammatical patterns are distributed systematically differently according to the communicative purposes and situations of use of the texts under study (e.g., Biber 1988; Biber et al. 1999). Pedagogical approaches to language description, however, have, if at all, tended to focus on linguistic variation across national (and more rarely regional and sociocultural) varieties of English only. This is highly problematic given that the few studies that have attempted to quantify both generic and geographic dimensions of variation have consistently found that text genre/register is a much more powerful predictor of variation than geography (e.g., Bohmann 2017). Similarly, textbook language evaluations have traditionally either considered textbook language as one register (e.g., Ljung 1990), thus disregarding major intra-textbook register variations, or focused solely on textbook dialogues (e.g., Mindt 1995b; Römer 2005). Aiming for a comprehensive lexico-grammatical analysis of Textbook English, this study accounts for the different registers featured in school EFL textbooks, in recognition of the fact that "[s]trong patterns in one register often represent only weak patterns in other registers, and, consequently, few descriptions of language are adequate for a language as a whole" (Barbieri & Eckhardt 2007: 325). As a result, the present study aims to explore the lexico-grammatical specificities of a range of typical textbook registers. To this end, the main textbook registers featured in school EFL textbooks were first identified. Before this procedure is described, a few definitions are in order.

Numerous attempts have been made to tease apart the often partly overlapping terms *genre, register* and *text type*. For many researchers, the use of the term *genre* or *register* is a matter of tradition or personal choice; both have been used to refer to text varieties associated with specific situations of use and communicative purposes (Egbert, Biber & Davies 2015). Biber & Conrad's (2019) framework, however, distinguishes between the two terms. *Genres* are text categories whose definitions are based on their conventional structures. In general, *genre* studies have tended to focus on socio-cultural aspects (Biber 2006: 11). *Registers*, on the other hand, are text categories that are defined according to their situational characteristics, such as their communicative purpose, the type of interaction and participants they involve, and topic (though the latter is controversial, see Lee 2001) (e.g., Biber 2006; Egbert, Biber & Davies 2015; Biber & Conrad 2019).[13] *Register* studies home in on the typical

[13] It is worth mentioning, however, that in pedagogically-motivated systemic functional linguistics (SFL), the term *genre* has frequently been used to refer to what most other SFL research refers to as

lexico-grammatical features of particular registers, thus revealing the systematic use of specific features in particular contexts of use. Finally, the term *text type* refers to text varieties which are initially defined according to similarities in linguistic form, i.e., in the co-occurrences of lexico-grammatical features. In other words, the terms *genre, register* and *text type* refer to different, external or internal, yet complementary, perspectives on text varieties (for a more detailed discussion, see Lee 2001). For the purposes of this study, the term *register* is preferred because the different text varieties found in the textbooks are initially distinguished according to their situational features, before being functionally analysed on the basis of their specific linguistic features. Put differently, it is assumed that a register's defining lexico-grammatical features serve a functional purpose within a particular situational context of use. For instance, it is known that most face-to-face conversations call for the frequent use of first and second personal pronouns (Biber 1988).

As mentioned earlier, textbook corpus research has largely evaded the question of register in textbook language. This is possibly due to the emphasis placed on the text unit in text-linguistic research that has traditionally been based on the assumption that "texts are nested within registers, but the opposite is not true: registers are not nested within texts" (Egbert & Mahlberg 2020: 75). In an article entitled 'Fiction – one register or two?', Egbert & Mahlberg (2020) break away from this tradition by analysing the linguistic variation within novels whilst distinguishing between passages of narration and fictional speech in fictional writing. As a result, coherent texts were divided into text segments that, while situationally different, remain contextually interdependent. To a certain extent, subdividing textbooks into texts of different registers may also be interpreted as dividing whole, coherent texts into text segments. In actual fact, defining text units, as required in text-linguistic corpus approaches where each text represents one observation (Biber et al. 2016: 357), is particularly tricky when it comes to textbooks. Indeed, typical school textbooks offer a range of plausible units of observation: the textbook series, textbook volume, chapter/unit, subchapter/unit, right down to the individual text (Le Foll 2020c). This study applies the smallest of these units where any one exercise, reading passage, explanation, instruction, dialogue or transcript corresponds to one observation. At the same time, each text observation is nested within one of the 43 textbook volumes and nine textbook series (see Table 5).

Text subdivision and annotation was performed manually. As part of this process, each text was manually annotated for register. First, however, it was necessary to identify meaningful register categories for all the texts featured in the 43 textbooks of the corpus. To the author's best knowledge, this had not been attempted in this form before. The most detailed textbook mark-up scheme the author is aware of is

register (this is particularly true of the Sydney School, e.g., Martin & Rose 2008; Martin 2009; Rose & Martin 2012).

that of the TeMa project (Meunier & Gouverneur 2009). The TeMa corpus consists of 32 English for General Purposes (EGP) textbook volumes. Both the coursebooks and workbooks of each textbook are subdivided into four subcorpora: texts, transcription of the tape scripts, vocabulary exercises and the guidelines to these exercises (Meunier & Gouverneur 2009: 7). Frequently, however, textbook instructional language has either been annotated for separate analysis or entirely removed from textbook corpora. For instance, in an exploration of lexical clusters in EAP textbooks, Wood (2010) organised his textbook corpus into two subcorpora: one containing the main textual elements of textbooks, and the other capturing the instructional material. He also pruned the raw text data of titles, headings, tables of content, prefaces, etc., thus obtaining a textbook corpus of approximately 580,000 tokens, of which 68% consisted of instructional material. In a different study comparing business and engineering textbooks with EAP textbooks, Wood & Appel (2014) acknowledged the difficulty arising from the wide range of registers found in textbooks. They therefore removed all instructional language from their business and engineering textbooks and annotated their EAP textbook corpus so as to create two subcorpora: one for the reading texts, and a second for all "instructional language" including vocabulary and comprehension exercises (which accounted for ca. 42% of the total word count).

This study aims to account for, among other factors, register-based variation in modelling Textbook English so that a simple division of (reading/listening) text vs. instructional language is not satisfactory. To determine suitable textbook text register categories, the following cyclical categorisation process (inspired by Mayring 2010: 84–85; Kuckartz 2014: 43–44) was applied:

1. Define the aim of the categorisation according to the research questions.
2. Set selection criteria according to the research aim and research questions.
3. Set appropriate abstraction levels for the register categories.
4. First cycle coding: Begin annotating one textbook from each series.
5. Proceed with category differentiation and subsumption.
6. End first cycle coding and verify all categories for duplication and/or incoherence in abstraction levels.
7. If necessary, re-define categories.
8. Second cycle coding: Re-annotate the textbooks that have already been annotated using the up-dated categorisation system.
9. Annotate the remaining textbooks.
10. Spot check the files for any annotation errors.

In order to reduce the manual annotation workload as well as the risk of inattention errors, short scripts were inputted into Keyboard Maestro 7.3.1 which enabled the two annotators to simply highlight each text in the raw textbook files and then press two keys for the shortcut corresponding to the correct register category (e.g., `cmd + I` for 'Instructional') for the section to be automatically annotated with the appropriate

XML syntax, e.g., adding `<div type="instructional">` before the highlighted passage and `</div>` at the end (see Le Foll 2020c for details and video demonstration). The final annotated version of the TEC files contains over 52,000 individual `div type` tags.

(1) `<div type="instructional"> Try to guess what each piece of information refers to. </div><div type="individual words or sentences">Olivia Timothy Liverpool London nd of March Fiona Loudon 4 Ella Steven Spielberg Casino Royale going to the theatre </div><div type="instructional"> Then listen to him speak about his life and check if you were right. </div>`

Although the annotation process was undertaken directly in the raw text files, it was frequently necessary to refer to the PDF versions of the textbooks in order to decide on the appropriate register. This manual register annotation process was very time-consuming; however, it also served to thoroughly check the OCR process across all the textbooks. This was important because the use of original fonts and the frequently complex formatting of blocks of texts on individual pages of the textbooks meant that it was often necessary to correct OCR mistakes and in some cases re-type or re-organise sections of the processed text files. In other words, the register annotation process also contributed to the cleaning of the raw data. In the rare cases where text was simply not retrievable (i.e., due to poor scanning), it was not annotated and therefore not included in the TEC so as not to compromise future analyses.

The register annotation was carried out by the author and a student research assistant. The reliability of the annotation scheme and method was tested by having both coders blind-annotate three full textbook volumes and comparing the results. Inter-rater agreement rate was found to be satisfactorily high (96.65%). The only notable difficulty consisted in distinguishing between individual sentences and isolated words/phrases; hence these two categories were merged into one in the final annotation scheme.

Ultimately, the cyclical categorisation process outlined above led to the creation of eight textbook register categories (see Table 6 and Fig. 2): Conversation, Informative writing, Fiction, Personal correspondence (letters, diary entries, social media posts, and e-mails), Instructional (instructions and explanations), Poetry (songs and poems), Other texts (timetables, shopping lists, etc.) and Individual words or sentences. Tags were also added to identify textbook passages in languages other than English (i.e., explanations or translations in the students' L1/school language). In the following, each register category and its rationale are briefly outlined.

Table 6: Number of words per textbook register categories in the TEC (as calculated by Sketch Engine Kilgarriff et al. 2014)

Register	Words
Conversation (spoken)	508,370
Fiction (narrative)	253,836
Individual words or sentences	913,331
Informative texts	302,739
Instructional texts	591,743
Personal correspondence (personal)	67,050
Poetry & rhyme (poetry)	26,174
Other texts in English (other)	14,379
Non-English texts (foreign)	346,336
Total	**3,023,958**

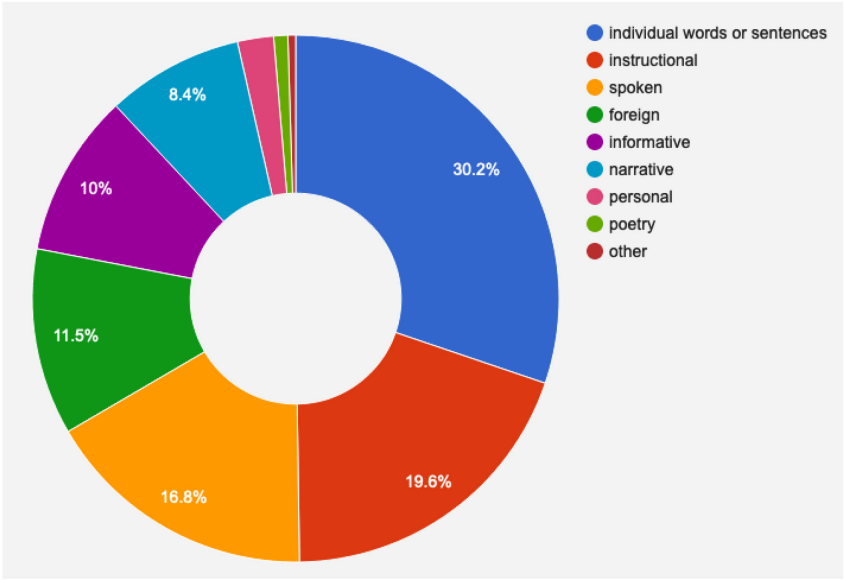


Fig. 2: Proportion of tokens in each register category on the TEC

In Chapter 2, we saw that some Textbook English studies have focused exclusively on the dialogues and other spoken-like language featured in EFL textbooks (e.g., Mindt 1992; Römer 2005). In this study, all dialogues, scripts and transcripts of the audio and video materials accompanying the textbooks purporting to be unscripted were annotated as one register labelled 'Conversation', e.g.:

(2) Nice of you to let us come to your barbie, Mike.
No worries. Great you're here. - Hey Cam! Come and meet a couple of new mates. They're staying at the hostel. Hey, how're you doing?
Hi. I'm Tanya. Nice to meet you.
You're a Kiwi, right? From your accent? And you're, let me guess ... American? No, I'm from Israel. Moshe.
OK, cool. You know Mike can trace his ancestors right back to the first British convicts in Australia?
Come on, Cam. Not that joke again! The story is: My ancestor Bill was walking down the road when a man bumped into him. The man was being chased by a police officer because he'd stolen a gold necklace from a

jewellery shop. When the man ran off, the police officer stopped Bill and found the gold necklace in his pocket. Then Bill was arrested and given a sentence of 20 years in Australia.
OK, he was just a victim. But many of the convicts were real criminals. [...] <TEC: New Green Line 5>[14]

The 'Informative' register tag was used to annotate factual articles, newspaper-like writing, reports (including scripted oral reports featured in video and audio materials) and texts from informative websites, e.g.:

(3) English is an official language in over seventy-five countries in the world.
More than two billion people speak English. Fifty-four English-speaking countries are members of the Commonwealth of Nations, an association of independent countries. Queen Elizabeth II is head of the Commonwealth.
31 % (percent) of the world's population live in the Commonwealth.
Six people out of ten in the United Kingdom have a relative in a Commonwealth country. <TEC: Hi There 5e>

'Fiction', or narrative, texts in lower secondary school EFL textbooks are mostly found in the form of short stories and extracts of novels, e.g., (4). To keep the integrity of the texts, direct speech passages in these narrative texts were not annotated separately (for a different approach, see Mindt 1995a: 7).

(4) With my backpack in my hands, I stepped off the train onto the crowded platform. It was 7:30 in the evening. People were hurrying home. A mother and her two young children were sitting on a bench. The mother was talking to the boy, but he wasn't looking at her. The girl was singing quietly and playing with a toy. Around them, travellers were shouting greetings, waving goodbye, carrying heavy bags or running to catch trains. A very tall man was standing completely still near the exit. Why was he wearing summer clothes in this weather? And why was he looking straight at me? […] <TEC: Solutions Pre-intermediate>

Singapore Wala (2013: 134) describes textbooks' "narratorial voice […] as the formal and most powerful means of structuring the relationship between coursebook and learner". As such, the coursebook narrator is responsible for informing, instructing learners to carry out learning tasks and asking questions to seek information in order for learners to meet their learning objectives. Here, the textbooks' "narrative voice" was annotated as 'Instructional', regardless of the exact function of the narrative voice at any point in time, e.g., (5). As has been shown by Wood (2010) and Wood & Appel (2014), this register is likely to make up a large proportion of school EFL textbooks and since textbook metalanguage is a major source of linguistic input for EFL learners (Kim & Hall 2002), it seemed essential to include it in this comprehensive exploration of Textbook English.

[14] Throughout this thesis, references to corpus excerpts are provided in angle brackets. The corpus abbreviation is followed by a colon and then a file identifier. In the case of the TEC, each corpus file corresponds to a textbook, hence the name of the textbook volume is printed.

(5) Plymouth, my hometown
a) In the film, the girl shows us her hometown. Watch the film. What did she show us? Choose A or B.
b) Watch the film again.
What other things and places can you see in the film?
Make a list. <TEC: Access 1>

The register category 'Personal correspondence' includes diary and blog entries, personal e-mails and letters, e.g., (6). Since formal letters and e-mails may serve somewhat different communicative purpose, they were originally allocated to a separate category. However, as lower secondary school textbooks contain very few such texts, they were later relegated to the category 'Others' as part of the second annotation cycle.

(6) Ally McKoene > WestHigh Bros
December 1 near University Heights, IA via mobile
Your best feature is definitely your kindness and I'm sure everyone else agrees! You have tons of kindness in your heart and your compliments can light up anyone's face. You guys are some of the kindest people I've met and I'm so glad that you guys do what you do. Your compliments can make anyone's day :) keep it up!
Like - Comment
nh [OCR error: Facebook-style thumbs up symbol] West High Bros likes this. <TEC: New Mission 2e>

Though not included in the original annotation scheme, the first cycle of the categorisation process revealed that songs, poems and rhymes feature heavily in lower secondary school EFL textbooks, thus justifying the need for a separate 'Poetry' register category, e.g.:

(7) School friends
Welcome to my school!
Welcome to my school!
Come in, and be cool!
Good morning, you can all sit down!
I'm Mister Parker
Yes, I'm your teacher
Good morning, Good morning Sir!
Can you repeat? I think I don't know I don't understand... Can I open the window?
Can you come to the board? Can you write the date? Yes sir! Yes sir!
It's a piece of cake!
Welcome to my school!
Let's all be cheerful!
Take your pens and write this down I'm Mister Parker!
Listen to your teacher!
Yes Sir! No Sir! Thank you Sir! <TEC: Join the Team 6e>

Finally, the register tag "Individual words or sentences" was used to label example words, phrases or sentences designed to illustrate particular lexico-grammatical phenomena, as well as the contextless words, phrases and sentences featured in

exercises, e.g., (8). Note that glossaries found at the back of many coursebooks were not included in the TEC.

(8) a) You are in a home, not a hotel!
b) Questions are better than mistakes.
c) It's important to be polite.
d) Your host family will get worried.
e) It can be very tasty!
f) Maybe there is a queue outside! <TEC: Green Line 2>

In addition, the manual annotation process revealed that the textbook series differ considerably in their use of students' L1/school language to provide instructions and explain grammatical points. As a result, it was thought worthwhile to keep track of the use of a language other than English across the various textbooks series and different proficiency levels. All extended passages in languages other than English were thus annotated with the register tag 'Foreign', e.g., (9). Note that, in practice, these passages are almost exclusively either in French or in German since the Spanish subcorpus of the TEC only includes series from Anglo-Saxon publishers with the international market in mind (see 3.3.1.1).

(9) Regarde la légende du document. a. Identifie la date. b. Cherche des informations à propos de l'artiste sur internet. B c. Lis le titre, regarde le tableau et devine le sens du mot *shiner*. <TEC: Hi There 3e>

In the following sections and chapters, the register subcorpora of the TEC are frequently referred to as separate entities. To differentiate the general reference to typical texts of a particular register found in textbooks from, specifically, a register subcorpus of the TEC, the word 'Textbook' and the first word of the register category are capitalised in Table 6 are used to denote the latter, i.e., 'Textbook Conversation' and 'Textbook Personal' refer to the Conversation and the Personal correspondence subcorpora of the TEC, respectively.

3.3.2 The reference corpora

As has been noted in 2.3, when analysing the frequency, use and function of individual lexico-grammatical features of Textbook English, a realistic reference benchmark is of the upmost importance. It has been argued that target learner language is, for instance, unlikely to resemble professional journalistic writing, and that such comparisons are thus unhelpful indicators when evaluating the lexico-grammatical content of Textbook English. It is also an argument advanced by Harwood (2005), in his criticism of studies that compare ESL textbooks to corpora. Striking a similar tone, Miller (2011: 34) rightly argues that "we must carefully consider measures (e.g., comparison corpora) upon which we are gauging our evaluation so that conclusions drawn are indeed fair and useful". Thus, it is not necessarily meaningful to compare learner language to professional native-speaker writers and radio presenters, using such general corpora as the British National Corpus (Burnard 2007), since this is not what secondary school students are expected to aspire to. On the basis of this claim,

Tono (2004) somewhat counterintuitively argues that Textbook English itself may be a useful benchmark to compare learner language to. However, if we are concerned with the authenticity and relevance of the language input that EFL learners receive in the EFL classroom, this approach is evidently cyclical. It certainly runs the risk of learners achieving only "textbook proficiency" (to borrow a term from Dörnyei, Durow & Zahran 2004: 87) rather than the language competences required beyond the EFL classroom.

That said, setting out to compile a more realistic target learner language reference corpus is not without its issues. In the field of tertiary-level learner English, however, such a project has already been undertaken. Indeed, when Biber et al. (2002; 2004) embarked on the TOEFL 2000 Spoken and Writing Academic Language Project (T2K-SWAL), their initial motivation grew out of the need to create "an external standard to evaluate the representativeness of ESL/EFL materials" (Biber 2006: 20). The T2K-SWAL corpus emerged from this large-scale project. In an initial phase, it served as a basis to identify salient lexico-grammatical features and patterns used across a range of university registers and academic disciplines (for a detailed report on the project design, corpus compilation and its subsequent linguistic analysis, see Biber et al. 2004). Drawing on these results, several diagnostic tools were developed to evaluate the extent to which a text can be considered representative of its target reference register.

The difficulty of finding an appropriate benchmark corpus for school EFL textbooks is compounded by the fact that, as shown in 3.3.1.4, textbooks comprise several, quite distinct registers, as well as text passages that consist of individual, contextless phrases and sentences of no identifiable register category. Instead of attempting to compile a single reference corpus for the TEC, it was decided to focus on three of the major registers identified in school EFL textbooks: Conversation, Fiction and Informative texts. The choice of the Spoken BNC2014 as the reference corpus for the conversation transcripts and conversation-like dialogues featured in the TEC is justified in 3.3.2.1. Section 3.3.2.2 explains how a corpus of modern fiction literature aimed at children, teenagers and young adults was compiled to match the Fiction subcorpus of the TEC. Finally, Section 3.3.2.3 outlines the rationale and design of the Informative Texts for Teen Corpus, compiled from web data as a reference corpus for the Informative texts featured in the EFL textbooks of the TEC.

3.3.2.1 The Spoken BNC2014

A number of earlier Textbook English studies that focused on the spoken or spoken-like passages of EFL textbooks relied on the demographically-sampled section of the British National Corpus 1994 (hereafter BNC1994; 4.2 million words of transcribed conversation recorded in the early 1990s in the UK) as a reference corpus for comparisons between textbook vs. authentic language use (e.g., Römer 2005). In the

present study, Textbook Conversation is compared to the equivalent component of the latest version of the British National Corpus, the Spoken BNC2014 (Love et al. 2017; 2018). The new Spoken BNC2014 is an 11.4-million-word corpus of orthographically transcribed conversations among L1 speakers in the UK (covering a range of self-reported regional dialects). The recordings were made by the speakers themselves using their own smartphones between 2012 and 2016 (Love, Hawtin & Hardie 2018: 4–5). In total, the corpus features 668 speakers in a total of 1,251 recordings (Love, Hawtin & Hardie 2018: 1).

The Spoken BNC2014 was chosen because, to date, it is the largest publicly accessible corpus of contemporary spoken English and one of the very few to reflect unscripted, informal conversation on everyday topics, which was identified as a key register in Textbook English (see 3.3.1.4). In choosing the Spoken BNC2014 as a reference corpus for Textbook Conversation, two additional choices were made: the choice of a native English variety as a reference for Textbook Conversation on the one hand, and the choice of British English over US-American or any other L1 English variety, on the other. The thoughts and reflections that motivated these two choices are explained in the following brief excursus.

On the use of L1 norms in English language teaching

Traditionally, EFL instruction in Europe has largely relied on British English norms (see, e.g., Bieswanger 2012; Forsberg, Mohr & Jansen 2019; Gilquin 2018). However, over the past few decades, "the whole notion of *nativeness* has become murky, if not downright controversial" (Moyer 2013: 91, emphasis original). In ESL contexts, the hegemonic, colonial implications that lie at the heart of many native vs. non-native distinctions are no longer tenable. But even in EFL instructional contexts, such as those found in most of continental Europe, the relevance of native speaker models has increasingly been questioned as a result of the ever-growing use of English as a *lingua franca* (ELF) and English as an International Language (EIL) in the business world and beyond (e.g., Jenkins 1998; 2000; 2003; Gnutzmann 1999; Gnutzmann & Intemann 2008; Prodromou 1992). If 21st century English teachers aim to follow a communicative language teaching approach, they will evidently need to equip their students with the (socio-)linguistic and pragmatic knowledge to communicate with both native and non-native speakers of English from various regional and socio-cultural backgrounds (Bieswanger 2012: 27). This is why many have called for a definition of "authentic" foreign language exposure that recognises "the reality of language use which learners will encounter outside and after their course[s]" (Tomlinson 2013b: 476). In the case of EFL teaching in continental European schools, this post-instruction context will likely involve interacting with native and non-native speakers of English in both professional and personal ELF contexts. Thus, there are many convincing arguments for deciding against the exclusive reliance on native-speaker norms in EFL instruction.

Correspondingly, Mauranen (e.g., 2003; 2004a; 2004b; 2010) advocates for the use of ELF reference corpora for pedagogic purposes (such as the English as a Lingua Franca in Academic Settings Corpus [ELFA]) which she claims contains "good international English spoken in academic and professional contexts" (Mauranen 2004a: 207). However, compiling such a reference corpus is not without its issues. It goes without saying that selecting "good speakers" (see also Prodromou 2003 on the notion of "successful users of English") to include in such a corpus will inevitably involve some subjective and/or normative judgements which will, themselves, be based on the researchers' own (naturally biased) norms and standards (cf. Mauranen 2004a: 207). This problem is, of course, not limited to the compilation of non-native language corpora. It can easily be argued that many native speakers of English are, in fact, not particularly eloquent speakers (cf. McCarthy & Carter 2001: 339). And when it comes to written registers, even fewer L1 users are expert writers of the kinds of professionally written texts featured in most general L1 corpora such as the BNCs or COCA. Suffice to say that the idea of capturing the language of proficient or expert language users regardless of their native-speaker status is – no matter how theoretically or pedagogically meaningfully – rather difficult in practice.

Whilst recognising the need to raise awareness of and embrace the diversity of plural native, non-native, standard, and non-standard English varieties in use around the globe, some English education scholars argue that EFL teachers and their students nevertheless need "a standard for pedagogical consistency" (Moyer 2013: 92). Thus, although native speaker norms are no longer explicitly mentioned in European secondary school EFL curricula (see 1.2), they nevertheless remain the most practical and reliable way of evaluating whether students are exposed to and themselves produce "authentic" and "correct" English – to quote two adjectives still frequently found in secondary school EFL curricula (see 1.2). On the other hand, staunch advocates of communicative foreign language teaching approaches argue that such a strict understanding of "authentic" and "correct" language use writes off too many unidiomatic learner usage cases as "inauthentic" and "incorrect" when they are, in fact, frequent in ELF contexts and do not hinder communication among native or non-native speakers of English. Mauranen (2004a: 208) illustrates this with an example from Altenberg and Granger's (2001) investigation of L2 speakers' use of collocations with the verb MAKE: she concludes that many, when compared to L1 use, so-called "collocation errors" are actually irrelevant for daily communicative needs and should therefore not be highlighted as deviant.

At the opposite end of the spectrum, Siepmann et al. (2011: 4) remind critics of native speaker norms in (advanced) foreign language teaching of "the age-old insight that the lower you set your sights, the less you will ultimately achieve". Returning to the example of collocates of MAKE mentioned above, they convincingly argue:

> the word combination 'make a claim' could theoretically mean 'invent a claim', but there is a common-sense convention which assigns to it the meaning 'utter an assertion'. There is, of course, nothing that prevents foreign-born writers [in the sense of: L2 English users] from using 'make a claim' creatively to mean 'invent a claim'; the snag is that their (unidiomatic) use of the word combination is certain to be misinterpreted by both native and non-native speakers of English (Siepmann et al. 2011: 4).

Going further, Siepmann et al. (2011: 4) warn that "[o]nce you start turning a blind eye to [standard L1 norm infringements], it is difficult to say where to draw the line" – an argument that echoes that of the difficulty of reliably distinguishing between (very) proficient and so-called "non-proficient" speakers of English (regardless of their native-speaker status). Gilmore (2007: 106) adds that taking the production of even highly proficient (however this may be defined) L2 speakers as the reference norm runs "the risk of providing learners with 'dumbed down' models of English which, although perhaps meeting their transactional needs, fail to illustrate the true expressive potential of the language".

A further argument in favour of L1 norms can be found in attitudinal research, which has repeatedly shown that both EFL teachers and, crucially, EFL learners still largely aim for native-speaker norms in spite of the generalisation of ELF/EIL (e.g., Edwards 2016; Forsberg, Mohr & Jansen 2019; Mohr, Jansen & Forsberg 2019; Scales et al. 2006; Timmis 2003). Even when their foreseen use of English is more likely to be with other non-native speakers in international ELF contexts, English L2 learners nevertheless claim to aim for norms aligned with those of Inner Circle English varieties, i.e., from regions with largely monolingual English-speaking populations (Kachru & Smith 2008). Of course, there is no denying that students' opinions will be shaped by all kinds of societal pressures including their teachers' (perceived or presumed) preferences. Nonetheless, the results of such attitudinal research should be taken into consideration if we are to attempt to break away from the customary paternalistic approach that tends to cast aside the opinions and wishes of learners as irrelevant (Wain 1992: 24).

In addition, some ELF agendas (as laid out, for instance, in Jenkins 2000; Seidlhofer 2001) risk dissociating learning to communicate in English from the (arguably also highly relevant) sociocultural contexts of the language – a critical aspect repeatedly highlighted as "intercultural (communicative) competence" in the EFL curricula of the educational authorities of France, Germany and, to a lesser extent, Spain (see, e.g., Conseil supérieur des programmes 2015; Consejería de Educación, Juventud y Deporte de Madrid 2015; Kultusministerkonferenz 2003; Kultusministerkonferenz 2012), as well as the CEFR (Council of Europe 2001; 2020) on which European national and regional school curricula are increasingly based on.

Bearing these three factors in mind – the pedagogical need for (at least some) consistency, respecting learners' wishes (though there is no doubt that these are

influenced by those of their peers, teachers, parents and society as a whole) and the curricular requirements to also teach English for intercultural competence, in addition to the pragmatic considerations mentioned above (in particular: what constitutes a "good/proficient" ELF speaker?), no attempt was made to create ELF reference corpora for this project. This is not to say that students' language production cannot or should never be assessed against an ELF norm (however hard that may be to define!). Indeed, the present study of Textbook English focuses on the evaluation of students' language *input* rather than *output* and this input is expected to be largely shaped by 'standard' L1 usage.

Given that the three educational systems represented in the TEC are situated in Europe, the choice of British English over the other dominating L1 English variety in EFL instructional contexts – American English – seemed most natural. As mentioned at the beginning of this section, British English has been, and continues to be, the most commonly used target English variety in Europe, even though the relevant national/regional curricula no longer explicitly refer to a single British English norm (see 1.2). In the Netherlands, a large survey concluded that British English remains the English model of choice for over half of the Dutch population, whilst just 15% claimed to aim for an American English norm (Edwards 2016: 81). Similarly, the results of an attitudinal study conducted in Spain suggest that learners prefer Standard British English (RP) rather than American English and aim to emulate this variety themselves (Carrie 2017). In practice, all the textbook series featured in the TEC focus on England, both in terms of language use (most noticeably, of course, in the pronunciation in the audio and video materials – though this is not the focus of this thesis) and cultural contextualisation. That said, individual activities and textbook units do attempt to feature other L1 speaker varieties, most notably US-American[15], Australian, Irish, and South African Englishes (see Scheiwe in preparation on how realistic the portrayals of such accents, usually produced by British actors, are in German textbooks). Across all the textbooks of the TEC, however, British English is clearly the target norm since it is deviations from British English that are indicated; for example, in most textbook glossaries, the terms *movie* and *cellphone* are followed by "American English", "AmE" or similar.

As its name suggests, the Spoken BNC2014, which was chosen as the reference corpus for Textbook Conversation, is most representative of British English – though it covers a range of regional dialects. The corpus features 566 English L1 speakers from all over the UK, 17 from outside the UK, and this optional speaker metadata is unavailable for 88 speakers (who, together, account for about 10% of the total word count) (Love, Hawtin & Hardie 2018: 24). The remaining two reference corpora, Youth Fiction and Info Teens, are also biased towards British English but

[15] In fact, in the G8 [*Gymnasium* in eight years] system in Germany, Year 8 is devoted to US-American English and culture.

considerably less so than the reference corpus for spoken English. Thus, 55% of the novels of the Youth Fiction corpus are by British authors, 31% by US-American authors, and the remaining 14% by authors from eleven different countries. The reference corpus for the informative texts of the TEC mostly contains texts of unknown authorship. However, the web domains from which they were sourced are principally from the UK and the USA, with a smaller percentage of texts from Australia and New Zealand. Hence, both these reference corpora can also be said to largely represent L1 English usage. Details of the composition of these corpora can be found in 3.3.2.2 and 3.3.2.3.

Processing of the Spoken BNC2014

The Spoken BNC2014 is richly annotated with detailed metadata on the speakers and the context of each conversation. In the untagged XML version of the corpus, which served as the basis for the preparation of the version used in the present thesis, the metadata is listed in the files' headers. The analyses carried out in the context of this study do not take this metadata into consideration; hence the headers were removed. The corpus also includes numerous other metatags, e.g., for paralinguistic sounds, pauses and overlaps. These were also removed. Table 7 summarises the regular expressions featured in the R script which was used to pre-process the untagged XML files of the Spoken BNC2014 (see also Online Appendix 3.3). Many of these pre-processing steps were necessary because the data contributed to the Spoken BNC2014 has been fully anonymised and therefore contains many tags of anonymised words and phrases. These tags were replaced with placeholders designed to ensure that the POS-tagger and dependency parser used to further process the corpus would correctly label them for word class and function. In addition, truncated words, which are rarely correctly identified by lemmatisers and POS-taggers, were removed.

As opposed to the BNC1994, the transcription scheme of the Spoken BNC2014 makes minimal use of punctuation and in fact only allows for question marks (Love, Hawtin & Hardie 2018: 37–38). Since automatic taggers and parsers are usually trained with punctuated texts, placeholder full stops were added at utterance boundaries that did not end in a question mark in order to reduce the potential for tagging errors resulting from a lack of punctuation. However, it is worth noting that these full stops markers were not used in any further linguistic analyses. Hereafter, the text files generated after these replacements are referred to as the "John & Jill version of the Spoken BNC2014 corpus" (see 'Replacement' column in Table 7 as to why this name was chosen).

Table 7: Summary of the regular expressions (regex) used to process the Spoken BNC2014 (see Online Appendix 3.3 for full script)

Description of tag	Search regex	Replacement
Header with full metadata	<header>.*</header>	[nothing]
Anonymised male name	<anon type=\"name\" nameType=\"m\"/>	John
Anonymised female name	<anon type=\"name\" nameType=\"f\"/>	Jill
Anonymised neutral name	<anon type=\"name\" nameType=\"n\"/>	Sam
Anonymised place	<anon type=\"place\"/>	IVYBRIDGE
Anonymised telephone number	<anon type=\"telephoneNumber\"/>	0123456789
Anonymised address	<anon type=\"address\"/>	ADDRESS
Anonymised e-mail address	<anon type=\"email\"/>	anonemail@email.com
End of utterance not immediately preceded by a question mark	(?<!\\?)</u>	.
Truncated word	<trunc>.{0,12}</trunc>	[nothing]
Anonymised financial details	<anon type=\"financialDetails\"/>	FINANCIAL DETAILS
Anonymised social media name	<anon type=\"socialMediaName\"/>	@SAM
Anonymised data of birth	anon type=\"dateOfBirth\"/>	DOB
Other anonymised personal information	<anon type=\"miscPersonalInfo\"/>	PERSONAL INFORMATION
All other remaining tags	<.*?>	[nothing]

3.3.2.2 The Youth Fiction corpus

Off-the-shelf corpora of English fiction exist, e.g., the English subcorpus of the PhraseoRom Corpus (Novakova & Siepmann 2020) and the US Novel Corpus (Chicago Text Lab 2020). In addition, most general English corpora include a literature subcorpus of extracts of novels, e.g., the British National Corpora (1994 and 2014). However, for a meaningful comparison of Textbook Fiction with authentic fiction texts to be possible, it was decided that both the communication purposes and the intended target audiences of the texts ought to be matched. Thompson & Sealey (2007) report a number of significant differences between adult and children fiction in the frequencies and contextual uses of the most frequent types, parts-of-speech, lexical verbs, 4-grams and POS-grams. Though less surprising, their additional conclusion that the most frequently occurring semantic categories found in adult and children fiction differ considerably is no less relevant to the present study. Children's literature appears to have a much stronger focus on animals and other living creatures, food, plants and communication (due to the prevalence of direct speech and speech acts). Thus, Thompson & Sealey (2007) highlight the strikingly different representations of world and self in children and adult fiction. Their results confirmed

the need for creating a dedicated Youth Fiction Corpus to be used as a comparison corpus for the Fiction subcorpus of the TEC: Textbook Fiction.[16]

Since the TEC consists of textbooks intended for ca. 11 to 16 years-olds, the aim was to compile a balanced and representative corpus of English-language fiction books suitable for children, teenagers and young adults. Unlike films, books are not usually explicitly labelled as being suitable or targeted at particular age groups; it was therefore necessary to find alternative selection criteria. In an attempt to achieve sample representativeness and balance, the books to be included in the reference Youth Fiction corpus were selected from the following seven lists. The lists were chosen to represent the choices made by respected British and US-American media, as well as those made by the wider internet community, as represented in the two dynamic user-contributed lists from goodreads.com and NPR.org.

- The List: 100 Best Children's Books of All Time (published on 8 January 2015) <http://time.com/100-best-childrens-books/>
- LIST: The 100 Best Young Adult Books of All Time (published on 8 January 2015) <http://time.com/100-best-young-adult-books/>
- The 100 best children's books of all time (published on 19 July 2018, accessed on 13 September 2018) <https://www.telegraph.co.uk/books/childrens-books/100-best-childrens-books-time>
- The Guardian Children's Fiction Prize Winners (from 2000–2016, accessed on 20 January 2019) <https://en.wikipedia.org/wiki/Guardian_Children%27s_Fiction_Prize>
- The School Reading List - Suggested reading books for primary and secondary aged children in the UK (Years 7 and 8) (by Jan Tolkien, last updated on 20 January 2019 when accessed on 31 January 2019) <https://schoolreadinglist.co.uk/category/reading-lists-for-ks3-pupils/>
- What Book Got You Hooked? (user-contributed list, accessed on 30 January 2018 with 9,003 contributors at the time) <https://www.goodreads.com/list/show/651.What_Book_Got_You_Hooked_>
- Your Favorites: 100 Best-Ever Teen Novels (user-contributed list with 75,220 contributors, published on 7 August 2012) <https://www.npr.org/2012/08/07/157795366/your-favorites-100-best-ever-teen-novels?t=1539242729260>

One major drawback of this approach is undoubtedly the highly subjective nature of such best-of lists; yet, in the absence of book sale numbers, they provided a useful starting point. Picture books clearly aimed at children younger than 10 years were excluded, as were translations of books originally in languages other than English.

[16] The use of the relevant subcorpora of the Oxford Children's Corpus (Banerji et al. 2013; Wild, Kilgarriff & Tugwell 2013) was also considered. However, the compilers of the corpus did not respond to the author's request for access for the purposes of this project.

The final selection from the lists was opportunist and entirely based on the immediate availability of the books[17] in digital format (Epub or PDF). In part, however, it can be assumed that the availability of the books in digital format is generally a testimony to their popularity. The digital books were subsequently converted to UTF-8 text using the same OCR software as for the TEC and automatically cleaned of unwanted characters and systematic OCR errors using regular expressions, as described in 3.3.1.2.

In total, 300 books were collected, amounting to over 20 million words. The majority of the novels in the corpus were written by British authors (166 books), a large proportion by US-American authors (92 books) and the rest from eleven other countries including Australia, India, and Ireland. The full list of works included in the corpus may be found in the Online Appendix 3.4. Due to the nature of such best-of lists, the median initial publication date is 1994 and just 26% of the books in the corpus were first published between 2007 and 2017.

The Youth Fiction corpus consists of four random samples of approximately 5,000 words (splitting was performed at sentence boundaries, hence the slightly varying word counts) extracted from each of the 300 books collected and processed for the corpus, except for three very short books, which were only sampled once in full. With a total of 1,191 Youth Fiction texts, this procedure resulted in a number of texts comparable to that of the Spoken BNC2014.

3.3.2.3 The Informative Texts for Teens Corpus (Info Teens)

Whilst corpora of children's or young adults' fiction do exist, it was clear from the outset that the reference corpus for the Informative subcorpus of the TEC, Textbook Informative, would have to be compiled specifically for this project. The aim was to find informative texts that are targeted at English-speaking teenagers of the kinds of topics typically featured in school EFL textbooks. To this end, a list of 20 quality informative websites for teenagers from various English-speaking countries was compiled.

The web scraping process was facilitated by Sketch Engine's (Kilgarriff et al. 2014) corpus-building tool, which relies on the WebBootCaT (Baroni et al. 2006), to create a text corpus from the list of selected websites. Sketch Engine automatically downloads the text materials from all the relevant webpages, removes non-text elements, boilerplates, and duplicates to produce (relatively) clean text files. Of the 20 websites, four were later discarded because they did not permit text scraping.

[17] Many thanks to the PhraseoRom team (in particular Johan Didier and Susanne Dyka) who kindly provided text versions of the titles already included in the PhraseoRom corpus.

Post-duplication removal, Sketch Engine retrieved and cleaned 17,014 files from the remaining 16 web domains. As shown on Fig. 3, these texts were not distributed evenly among the websites. The corpus as originally compiled by Sketch Engine was then downloaded as a single XML file without part-of-speech tagging and lemmas for further off-line processing.

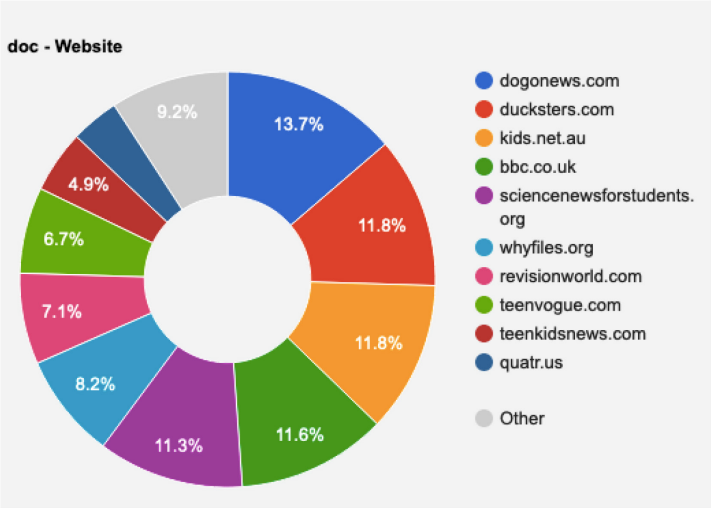


Fig. 3: Composition of the web corpus originally retrieved by Sketch Engine

Off-line processing was performed in python 3 (Van Rossum & Drake 2009). Regular expressions were employed to remove erroneous XML tags, non-UTF-8 characters, indices and tables of contents, any remaining boilerplates and adverts, texts containing language puzzles (e.g., crosswords), marking schemes and past exam papers (only in the texts from revisionworld.com), and user comments (especially in the texts extracted from dogonews.com, teenvogue.com and teen.wng.org). Then, the large XML file (totalling nearly one million lines!) was split into the individual texts of the corpus and saved as separate plain text files with filenames that incorporate relevant metadata on the web domain and title of the webpage, as available in the corresponding XML tags. This was achieved with the {beautifulsoup} library (Richardson 2015), which was found to cope relatively well with large, malformed XML files such as those compiled by Sketch Engine.

This procedure led to the creation of 10,104 individual text files, of which 4,895 were under 400 words and were therefore discarded since they were considered too short for any meaningful text-based analyses (see also 6.2.2). To achieve a more balanced corpus, a stratified sampling approach was followed: 100 texts from each web domain were randomly selected for inclusion in the corpus. For two domains, fewer than 100 texts longer than 400 words had been retrieved; for these, the full domain datasets were retained. The final selection thus consists of 1,414 text files (see Table 8) – a number comparable to both the total number of conversation files in the Spoken BNC2014 (see 3.3.2.1) and text samples in the Youth Fiction Corpus (see 3.3.2.2).

Table 8: Composition of the Informative Texts for Teen Corpus

Domain name	**Nb. texts**	**Nb. words**
bbc.co.uk/history	100	74,722
dogonews.com	100	60,762
ducksters.com	100	67,894
encyclopedia.kids.net.au	100	74,566
factmonster.com	100	60,395
historyforkids.net	100	71,955
quatr.us	100	62,254
revisionworld.com (GCSE only)	100	74,301
sciencekids.co.nz	100	57,097
sciencenewsforstudents.org	100	82,258
teen.wng.org	85	45,515
teenkidsnews.com	100	81,765
teenvogue.com	100	82,117
tweentribune.com	29	26,166
whyfiles.org	100	85,492
Total	1,414	1,007,259

4 Exploring the progressive in Textbook English

Are you swimming in the sea?
Yes, I am.
Are you playing golf in the jungle?
No, I'm not.
<TEC: Achievers Pre-intermediate>

4.1 Rationale and aims

This chapter[18] aims to investigate the representation of the progressive in two textbook registers: Conversation and Fiction. To this end, 3,940 concordance lines featuring verbs in the progressive from the Conversation and Fiction subcorpora of the Textbook English Corpus (TEC) are compared to the same number of progressive concordance lines from the two reference corpora of the same registers: the Spoken BNC2014 (see 3.3.2.1) and the Youth Fiction Corpus (see 3.3.2.2). First, a method is developed to approximate the frequency of the progressive construction in the two Textbook English subcorpora as compared to the two reference corpora. The second section focuses on contextual, morphosyntactic aspects of progressive constructions, such as contraction and negation, and their functions, including time reference and repeatedness. Finally, a third section explores the lexical and semantic associations of the progressive using collostructional methods and correspondence analysis.

4.1.1 EFL learners' use of the progressive

Despite its simple form, BE + V_{ing}, the English progressive aspect is fraught with pitfalls which EFL teachers will no doubt be familiar with. It has been claimed that it "constitutes one of the most basic and ubiquitous problems facing language teachers" (Williams 2002: 18, cited in Römer 2005: 172). Indeed, although the progressive is usually introduced in the first term of EFL instruction, the progressive vs. non-progressive alternation remains a stumbling block for many learners of English, even at the higher levels of proficiency (e.g., Hahn, Reich & Schmied 2000; Mindt 1997b; Rautionaho, Deshors & Meriläinen 2018; Westergren Axelsson & Hahn 2001; Wulff & Römer 2009). For many EFL learners, the progressive is tricky in that

[18] Preliminary results of the analyses reported in this chapter were presented at the ICAME40 in Neuchâtel, Switzerland. The author is grateful for the audience's valuable comments and suggestions which contributed to improving this chapter. The conference presentation also led to the publication of selected results from this chapter in: Le Foll, Elen. forthcoming. *"I'm putting some salt in my sandwich."* The use of the progressive in EFL textbook conversation. In Susanne Flach & Martin Hilpert (eds.), *Broadening the spectrum of corpus linguistics: New approaches to variability and change*. Amsterdam: John Benjamins.

it may not have a direct equivalent construction in their L1 (Römer 2005: 2). An added difficulty for EFL learners is undoubtedly that the construction fulfils a broad range of functions (see, e.g., Binnick 2020; Gut & Fuchs 2013: 245–246; Mair 2012). Moreover, numerous studies have shown that its diverse range of functions has seen considerable changes over the past decades (e.g., Mair & Hundt 1995; Smith 2005; Smitterberg 2005).

The progressive aspect has repeatedly been shown to be frequently over- and misused by EFL learners, especially those acquiring English in instructional settings. For instance, learner corpus studies have concluded that L1 Finnish, Swedish and German learners generally overuse the progressive as compared to English native (ENL) speakers (Virtanen 1997; Hundt & Vogel 2011; Westergren Axelsson & Hahn 2001; Römer 2005). More recent, fine-grained learner English studies have revealed how EFL learners' use of the progressive differs from ESL and ENL use in terms of frequency, range of semantic functions, choice of verb lemmas and genre-awareness, among other aspects (e.g., Fuchs & Werner 2018; Meriläinen 2018; Rautionaho & Deshors 2018; for comprehensive overviews of earlier studies, see also Römer 2005: chap. 3; Rautionaho 2014: chaps. 1–2).

We may ask whether these differences in the ENL, ESL and EFL speakers' use of the progressive can, at least partly, be attributed to EFL teaching practices. Concerning EFL learners' reported overuse of the progressive, Biber & Reppen (2002b: 203) claim that "[o]ne of the most widely held intuitions about language use among TESL professionals is the belief that progressive aspect is the unmarked choice in conversation." Whilst progressive constructions have been shown to be comparatively more frequent in spoken than in written English, corpus-informed grammars report that progressive are, in fact, rare across all varieties and registers, accounting for between some three and ten percent of all verb phrases (Biber et al. 1999: 461–475; Leech, Rayson & Wilson 2001: 124–125). At the same time, the use of the progressive in English appears to be generally on the rise (e.g., Hundt 2004; Mair & Hundt 1995; Smith 2005; Smith & Rayson 2007; see also Rautionaho 2014: 62–64 for a more recent overview of studies on the evolution of the frequency of the progressive). Hundt & Vogel (2011) concluded that both language contact and exo-normative influences are likely to contribute to the attested differences in progressive usage in ENL, ESL and EFL varieties. They suggest that instructed EFL learners are foremost confronted with the "grammaticality issues and semantic restrictions of the progressive, making them more likely to overuse the prototype of the construction and less likely to 'stretch' the progressive to new contexts, such as combinations with certain stative verbs or new aspectual uses" (Hundt & Vogel 2011: 160). All of this raises the question as to whether EFL grammars and textbooks have evolved to reflect current usage.

4.1.2 The progressive in EFL textbooks

It is beyond the scope of this chapter to provide even a brief summary of the extensive research that has hitherto been carried out on the patterns of use and functions of the progressive in various ENL, ESL and EFL varieties (see Römer [2005: Ch. 3] and Rautionaho [2014: Ch. 1 and 2] for comprehensive overviews). Instead, this section will focus on previous research on the use of the progressive in EFL textbooks for secondary education.

Compared to the many studies on English learners' use of the progressive, ELT materials have more rarely been the subject of investigation (see 2.2.2.3). Bland (1988) and Belli (2017) examined the treatment of stative verbs in the progressive in ESL/EFL pedagogical materials and concluded that it did not match authentic usage, whilst Biber & Reppen (2002b: 203) reported that EFL/ESL conversation textbooks overuse the progressive in their representations of spontaneous spoken language. However, to date, only Römer (2005) has provided a comprehensive, corpus-based analysis of the progressive in EFL textbooks.

As mentioned in Chapter 2, Römer (2005) examined the lexico-grammatical usage patterns of the progressive in a corpus of spoken-like language from two series of EFL textbooks used at lower secondary school level in Germany. She compared those to a reference corpus of spoken British English consisting of the spoken component of the British National Corpus (BNC1994) and the Bank of English. Some of the key findings were presented in 2.2.2.3. They included an under-representation of negated progressives and contracted forms of the auxiliary BE in textbooks as compared to naturally occurring speech, but an over-representation of progressives in question forms. Functionally, Römer (2005) observed a distinct underuse of progressives conveying repeated actions. Indeed, 90% textbook progressives refer to a single continuous event (Römer 2005: 6.2.2 and Römer 2010: 22–24) – a prototypical example being:

(10) *Robert: Well, no. I'm not listening to Radio 1. It's Radio Nottingham. I'm listening to my mum.* (extract from *Green Line New*, as cited in Römer 2010: 23).

Römer (2005) did not directly investigate which verb lemmas are frequently featured in the progressive, or the semantic domains of these verbs. However, her study frequently subdivides the analysis of specific morphosyntactic and functional features according to individual high-frequency verbs. Thus, for instance, she reports that present progressive tense forms with the verb lemmas ASK, GO, PLAY and STAY are over-represented in textbook dialogues (Römer 2005*:* 244–246). In contrast, she shows that these same verb lemmas are under-represented in the past progressive and present perfect progressive forms. Similarly, Römer (2005: 250–252) notes a general under-representation of negated progressives, in particular those featuring DO, GO,

HAVE and SAY, yet observes an overuse with ASK, GET, LISTEN and MAKE in textbook dialogues, as compared to her reference corpus of spoken British English. However, Römer's (2005) work does not attempt to tease apart the associations of individual verbs with the progressive, nor to compare these associations across textbook dialogues and real-life L1 conversations. An exception is made for phrasal/prepositional verbs in a section devoted to verb-preposition co-occurrences. Thus, in the progressive, the phrasemes AGREE WITH, DEAL WITH, RING UP and SET UP are found to be less frequent in textbook dialogues than in naturally occurring British English conversation, whereas COME BACK, GET AT, LOOK AT and WORK ON are reported to be more frequently observed in the progressive (Römer 2005: 249–250).

In conclusion to her Textbook English vs. authentic ENL English comparison, Römer (2005: 271–273) draws up a long list of the most striking disparities in the patterns of co-occurrences and function of the progressive in textbook and authentic dialogues. She concludes that school EFL "coursebooks [dialogues] provide a rather monolithic view of what progressives can express" (Römer 2005: 284). However, these results should be approached with caution since they are based on a relatively small corpus of pedagogical dialogues (108,000 words) stemming from just two series of textbooks used in German secondary schools. Effect sizes or confidence intervals are not reported for the many quantitative differences observed between textbook dialogues and natural conversation. In common with all Textbook English studies to date (see 3.1), the analyses do not enable us to tease apart which linguistic variable(s) contributed to any one reported difference in frequency of use (e.g., polarity or verb lemma, tense or time reference). Crucially, whilst the frequencies and distributions of progressive forms and functions in Römer's (2005) textbook dialogue corpus are frequently reported separately for the two textbook series that span five years of English tuition, no attempt is made to distinguish how representations of the progressive in school EFL textbooks may change as the target readers' language proficiency increases.

4.1.3 Aims and research questions

This chapter aims to investigate how the progressive construction is portrayed in two EFL textbook registers: Conversation and Fiction, as compared to the Spoken BNC2014 and the Youth Fiction Corpus, respectively. This differentiated look at the progressive across two different registers is crucial since the construction has been shown to be highly susceptible to mode, genre and register effects (e.g., Biber et al. 1999; Rautionaho & Deshors 2018; Smith 2005).

By investigating many of the linguistic parameters already studied by Römer (2005), it will be possible to compare the language of EFL textbooks currently in use in secondary schools in France, Germany and Spain to that of older German EFL textbooks. Following Römer (2005), the present investigation comprises both

contextual morphosyntactic features, such as tense forms, contractions and negation, and functional phenomena, such as time reference and repeatedness.

Since Römer's study did not attempt to assess the frequency of progressives, there is, to the best of the author's knowledge, currently no empirical data to back or refute the claim made by Biber & Reppen (2002b: 203) that progressives are over-represented in textbooks' representations of spoken English. This chapter therefore also investigates whether the progressive is over-represented in both the conversation and fiction texts of contemporary secondary school EFL textbooks.

In her concluding chapter on the pedagogical implications of her research, Römer (2005: 285) advocates for a shift towards a more lexical grammar of the progressive. Her analyses of verb-specific patterns reveal strong associations between progressive forms of individual verb lemmas and morphosyntactic features and functional phenomena with which they typically co-occur (Römer 2005: 285). Following a usage-based approach (see 1.4), we assume that lexical restrictions apply to grammar in general, and specifically here, to the progressive aspect. This chapter will therefore also apply a collostructional approach to investigating the lexical associations of the progressive in Textbook English. This will be achieved by comparing the results of pairs of Distinctive Collexeme Analyses (DCAs) of the lemmas associated with both progressive and non-progressive constructions in textbook dialogues and real-life conversation, on the one hand, and textbook fiction and non-textbook fiction, on the other. Although there have been successful multifactorial approaches to the study of progressives in different genres and ENL, ESL and EFL varieties (e.g., Deshors 2017; Fuchs & Gut 2015; Rautionaho, Deshors & Meriläinen 2018; Rautionaho & Deshors 2018; van Rooy 2006), so far, such models have not included verb lemma as an independent variable.[19] Excluding the variable lemma from such analyses, however, is potentially problematic both from a theoretical, usage-based approach, and a pedagogical perspective. Indeed, the aspect hypothesis, which is in turn motivated by the distributional bias hypothesis, postulates that:

> First and second language learners will initially be influenced by the inherent semantic aspect of verbs or predicates in the acquisition of tense and aspect markers associated with or affixed to these verbs (Andersen & Shirai 1994: 133).

In line with the predictions of the aspect hypothesis, Bardovi-Harlig (2012) advances that EFL learners first acquire the progressive aspect in the context of activity verbs and accomplishment. They later extend the construction to repetitive (iterative and habitual) actions, before beginning to use it in futurate and, finally, stative functions. Supporting the aspect hypothesis, Rautionaho & Deshors (2018) report that, in their final regression model of progressive use by ENL, ESL and EFL speakers, only the

[19] Rautionaho & Deshors (2018: 236) report that they did attempt to model lemma as a random effect variable but that such models failed to converge.

semantic domain variable does not contribute to any higher-level interactions. They thus conclude that "regardless of varieties, written genres, tense, Aktionsart [accomplishment, achievement, process, or stative], voice, etc., the semantics of a given lexical verb systematically influences writer's constructional choices" (Rautionaho & Deshors 2018: 238). Since lemma and semantic category are usually highly correlated, we might speculate that, had it been feasible to include them in such a model, individual verb lemmas may have played an even more important role than semantic domain in mapping the use of the progressive. In the context of this pedagogically-motivated study, it therefore seems imperative to also consider which specific verb lemmas are more strongly associated to the progressive (as opposed to the non-progressive) in EFL textbooks and real-life conversation.

Consequently, 4.3.4 zooms out from the broad contextual and functional aspects of the progressive to zoom in on the specific verb lemmas. Following the methodology proposed and tested by Deshors (2017) in an investigation of the use of the progressive in five varieties of World Englishes, it also explores the relationship between verb lemmas, the semantic domains of these verbs and the progressive construction. To this end, and following Deshors (2017), a series of Co-varying Collostructional Analyses (CovCAs) and a Correspondence Analysis (CA) are conducted. The final section takes an alternation stance on the lexical verbs featured in the progressives by comparing the results of two pairs of Distinctive Collexeme Analyses (DCAs) of the lemmas associated with both progressive and non-progressive constructions in the four (sub)corpora under study.

In sum, the present chapter seeks to address the following research questions:

1. Is the progressive over-represented in conversation and fiction texts of secondary school EFL textbooks?
2. What are the morphosyntactic contexts in which EFL learners encounter progressives in their textbooks? Are these contexts representative of those found in authentic ENL target-learner-language corpora of the same registers?
3. Which functions do progressives in Textbook English typically fulfil? Are these functions representative of the most common functions observed in the reference corpora?
4. Have the progressives featured in the dialogues of contemporary lower secondary school EFL textbooks become more representative of those typical of present-day authentic ENL conversation than was the case at the time of Römer's (2005) study?
5. Which lexical verbs and semantic domains are significantly associated with or repelled by the progressive in the conversation and fiction texts of EFL textbooks? How does this compare to naturally occurring ENL registers?
6. To what extent do the distinctive semantic domains of verbs in the progressive correlate with the registers Conversation and Fiction and with the varieties Textbook English vs. naturally occurring ENL?
7. Which of the differences noted in the use of progressives across Textbook English and the ENL reference corpora may be pedagogically well-founded? Which might

deprive learners of valuable exposure to frequent forms with particularly useful communicative functions?

4.2 Methodology

The corpus data analysed in this chapter were presented in Chapter 3. In the present chapter, two subcorpora of the TEC are examined: Textbook Conversation and Textbook Fiction (see 3.3.1.4). These are compared to the Spoken BNC2014 (see 3.3.2.1) and the Youth Fiction Corpus (see 3.3.2.2) respectively. The following sections focus on the methods applied in the analyses presented in this chapter. First, 4.2.1 explains how occurrences of the progressive were automatically extracted from the corpora and pre-processed. Next, 4.2.2 outlines how the extracted concordance lines were subsequently manually annotated for a number of morphosyntactic and functional variables. The statistical methods used to compare frequencies across the two corpora are discussed in 4.2.2.7, before the procedure followed to conduct the collostructional analyses of the lexical verbs attracted to or repelled by the progressive in textbook and naturally occurring English is explained in 4.2.3.

4.2.1 Extraction of progressive concordance lines

The aim of this extraction process was to retrieve all instances of the progressive from the two TEC subcorpora, Conversation and Fiction, and to subsequently extract, at random, the same number of progressives from each of the two reference corpora for comparison.

Many earlier corpus-based studies on the progressive did not make use of part-of-speech (POS) tagging to extract progressives, relying instead on a regular expression approach, i.e., combining a query for any token with the form *<.*ing>* with a labour-intensive manual disambiguation phase (Deshors 2017; Edwards 2014: 167; Rautionaho 2014: 73–74; Römer 2005: 50). Since English POS taggers are known to struggle with participle forms (see, e.g., Manning 2011), a non-POS-based query (*[word="[a-z]*ing"]*) was first run on the Textbook Conversation subcorpus and the Spoken BNC2014 to check for potential issues in using Sketch Engine's (Kilgarriff et al. 2014) POS tagging[20] as a basis for a query on progressives: for each corpus, a random sample of 200 POS-tagged concordance lines with tokens ending in *-ing* were thus manually checked. Of these, one concordance line (11) in the Textbook Conversation sample included a present participle form inaccurately tagged as a verb base form. Because of an end-of-line hyphen in the original textbook layout, *ting* was tokenised as a separate word. This is unfortunate but a subsequent corpus query confirmed that it was a one-off case across the entire TEC.

[20] English TreeTagger PoS tagset with Sketch Engine modifications, see https://www.sketchengine.eu/english-treetagger-pipeline-2)

(11) Thank you for helping me out, I'm get-**ting** a little tired of royal schedules [21] <TEC: Piece of Cake 3e>

Due to the unplanned and interactional nature of spontaneous conversation, the Spoken BNC2014 presents more of a challenge for POS taggers. Nonetheless, only two instances of poorly tagged present participle forms were found among the 200-concordance line sample. In (12) *being* is identified as an adjective because the expletive ending in *-ing* immediately before it is inaccurately identified as a present participle. Therefore, despite the erroneous automatic tagging, this occurrence of the progressive would still be picked up by a query relying on POS tags (albeit for the wrong lemma).

(12) I've never fucking met this girl and she was fucking **being** a dick was she actually shouting as well cos she was very quiet? <BNC2014: SAR5>

No plausible explanation for the second inaccurately tagged present participle (13) (tagged as a noun) could be found. In fact, the trigram *were you singing* occurs three times in the Spoken BNC2014 and is tagged correctly in the other two occurrences.

(13) what were you doing? were you **singing** or something? <BNC2014: SY8B>

At the same time, around a third of the concordance lines retrieved using this non-POS-based query were not present participles and were not tagged as such either. It thus becomes evident that, for a minimal drop in the recall rate, a POS-based query is likely to yield a 30% boost in precision (assuming that the concordance line samples were representative and that the POS tagging algorithm only assigns tokens ending in *-ing* to present participle tags). In the context of this study, it therefore seems reasonable to rely on Sketch Engine's POS tagging in order to considerably reduce the workload in the manual disambiguation phase. After some experimenting, the following Corpus Query Language (CQL) query (14) was chosen to extract progressives from the corpora stored in Sketch Engine.

(14) `[tag="VB.*"][tag!= "SENT|,"]{0,3}[tag="V.G" & word!="gonna"]`

The query first calls for all forms of the verb to BE. This includes the base form for modals with progressives, e.g., forms such as *will* + *be* + *-ing* (15), and the past participle form *been* for perfect progressives (16). Note that this CQL-based approach means that progressive constructions in which the auxiliary BE is omitted were not included in this investigation (unlike previous studies using an entirely manual filtering approach, e.g., Römer 2005, Rautionaho 2014). This is not deemed problematic since these make up less than 1% of progressive occurrences in spoken English (Römer 2005: 61).

(15) I probably can't get there till after five. Mum **will be working**. <TEC: Green Line New 5>

[21] Unless stated otherwise, all emphasis in example sentences added.

(16) I’**ve** never **been skiing**, but I’d love to go. <TEC: Solutions Pre-intermediate>

The CQL query (14) allows for the verb be to be optionally followed by one to three tokens of any POS except for end-of-sentence punctuation marks (SENT) or commas before the -ing verb form. The span of zero to three was chosen to allow for negation and up to two adverbs in declarative sentences, as well as for nominal phrases as subjects in questions, e.g., (17)–(18), and, particularly with the Spoken BNC2014 in mind, adverbs, discourse markers, and repeated words and/or hesitations placed between the auxiliary and the present participle, e.g., (19)–(20). Spot checks suggested that a larger span would not have led to a substantially improved recall rate, but clearly to a considerably lower precision rate.

(17) Are **those kids really** sleeping? < TEC: Access 1>

(18) yeah how’s **your university application** going? <BNC2014: SQPN>

(19) I’m **you know rapidly** losing patience with him <BNC2014: S2T6>

(20) and then as far as I could tell she was **essentially like** crashing at some guy’s place who like she’d just met <BNC2014: S3LE>

The word *gonna* was also excluded because it is mostly associated with *going to + infinitive* constructions expressing future actions and events, and this construction has been excluded from the present investigation (see 4.2.2.3).

4.2.2 Annotation

In order to reduce the manual annotation workload and thus decrease the rate of inattention errors, the concordance lines were automatically pre-annotated for aspects related to the forms of the progressives. This process is described in 4.2.2.1, before turning to the development, implementation and validation of the manual annotation scheme in the following subsections (4.2.2.2–4.2.2.6).

4.2.2.1 Automatic pre-annotation

When retrieving concordance lines using the CQL query described above (14), Sketch Engine gives the user the option of downloading the lines in a tabular format whereby the strings corresponding to the query results themselves are stored in one column, whilst their immediate left and right contexts are stored in two additional separate columns. Thus, it was possible to process only the strings retrieved by the CQL query, i.e., the middle of each KWIC concordance line, in order to pre-code each instance of a progressive for the following variables: a) the lemma of the verb in the progressive, b) whether or not the progressive includes a contracted form of the auxiliary BE, c) whether the verb phrase is negated and d) the tense form of the progressive. In the following, the R script written to this effect is briefly described. The script itself can be consulted in the Online Appendix 4.2).

The main verb lemma from each concordance line was determined by extracting the tokens with an *-ing* ending from the table column containing only the query results. This is easily achieved using a regular expression of the form *[:alpha:]*ing*. These tokens were then copied to a new variable and lemmatised using the R package {textstem} (Rinker 2018). Since the lemmatiser can then only lemmatise the verbs on the basis of these isolated present participle forms, this method is evidently not particularly accurate (e.g., *annoying* and *dying* were always lemmatised as the adjectives with the same form) but, in practice, this was not a problem because all the lines were, in any case, to be subsequently checked as part of the manual filtering and annotation phase (see 4.2.2.3).

A simple regular expression was also used to detect apostrophes since they likely point to contracted auxiliary verb forms. Similarly, the regular expression *grepl("n't/\\bnot\\b")* was used to capture negation in the captured progressive forms. This regular expression can detect both contracted and non-contracted negated forms and the word boundary metacharacter *\\b* was used to ensure that *not* was always an individual token, rather than detected as part of a longer word such as *noting* or *another.*

Automating the detection of tense was a little more complex, but since high accuracy was not the aim here, a similar approach with a series of regular expressions could be developed. The default value for the tense variable was set to 'present progressive' and occurrences of markers of past progressives *(\\bwas\\b/\\bwere\\b/ \\bwasn't\\b/\\bweren't\\b)* resulted in a past progressive coding, whereas the detection of *\\bbeen\\b* changed the default setting of the tense variable to 'perfect' progressives and *\\bbe\\b* to 'modal/infinitive' progressives. This last category, in particular, subsequently required considerable manual disambiguation since it included many concordance hits that were later excluded from this analysis of progressive verb phrases (see 4.2.2.3).

Evidently, such an approach does not pretend to yield highly reliable results. It should be stressed that the purpose of this automated pre-annotation phase was simply to reduce the total manual workload and, for this purpose, the script described above proved to be highly effective.

4.2.2.2 Semantic domains of progressive verbs

Following Deshors (2017), each verb occurring in the progressive was also coded according to the seven-class taxonomy of semantic domains developed by Biber et al. (1999: 360, emphasis in examples added):

- **Activity** verbs are used for events controlled by a volitional agent, e.g., BRING, GO, OPEN, RUN, TAKE.

- **Aspectual** verbs characterize the stage of progress of an event or activity, e.g., BEGIN, CONTINUE, KEEP, START, STOP.
- **Causative** verbs "indicate that some person or inanimate entity brings about a new state of affairs" e.g., CAUSE, HAVE, HELP, LET, MAKE.
- **Communication** verbs are a subcategory of activity verbs involving spoken and written communication, e.g., ASK, DESCRIBE, PROPOSE, SAY, WRITE.
- **Existence** verbs can be divided into verbs of existence or stance (e.g., BE, EXIST, LIVE, STAY, STAND) and relational verbs (e.g., APPEAR, BELONG, DEPEND, HOLD, RESEMBLE).
- **Mental** verbs and states verbs signal human experiences and fall into the categories of perception (SEE), cognition (THINK), decision (ACCEPT), mental effort or intent (AIM), and receipt of communication (READ).
- **Occurrence** verbs denote physical events that occur independently of volitional activity, e.g., BECOME, DIE, EMERGE, HAPPEN, INCREASE.

While chosen for reasons of comparability with previous research on the progressive (e.g., with Collins 2008; Deshors 2017; Edwards 2014; Rautionaho 2014; Smith 2005), this classification poses a few theoretical and practical problems. Conceptually, the main issue is that this framework does not include a distinct category for stative verbs, although this category evidently has the potential to be quite informative when investigating variation in the progressive construction (see also Deshors 2017: 266). In practice, many verbs are tricky to classify and thus the annotation process is subjective, time-consuming and, inevitably, prone to error. Deshors (2017) relied on data manually coded by Edwards (2014) as part of a larger project. Regrettably, it was not feasible to manually code over 7,000 progressive concordance lines for semantic domain in the context of this case study. Instead, a list of 345 verbs and their semantic domains mentioned in Biber (2006: 249) was merged with a list of 550 verbs belonging "fairly unambiguously to each of the seven categories" compiled by Alison Edwards (mentioned in Edwards 2014: 168fn, obtained via personal communication) which she used to pre-annotate concordance lines before manual disambiguation. Since the concordance lines were to be automatically automated for semantic domain, the category 'other/unclear' was added for all verbs which frequently belong to more than one semantic category. The few disagreements between Biber's and Edwards' lists usually concerned such polysemous verbs and, given that they cannot be reliably automatically annotated for semantic category, these were all re-classified as 'other/unclear'. The resulting agglomerated list was then manually expanded to include all remaining verbs found in the 7,880 progressive concordance lines under study (see Online Appendix 4.3). It was then possible to automatically match every concordance line to a semantic category for its verb, though it should be noted that many high-frequency verbs being highly polysemous, 26% of concordance lines were assigned the semantic category 'other/unclear'.

4.2.2.3 Manual identification of non-progressives

The manual annotation process began with the filtering of false positives, i.e., non-progressives that were unintentionally retrieved by the corpus query. Thus, a number of adjectives (e.g., *boring, interesting*), as well as fewer nouns (e.g., *living room*) that were erroneously tagged as present participles and occurred shortly after the verb BE were screened out. Furthermore, *BE going to + infinitive* constructions (21), non-finite verb phrases (22)–(24), *BE + present participle* (25) and unclear and/or incomplete constructions were excluded (see Smitterberg 2005; Rautionaho 2014 for similar procedures).

(21) Well, I**'m going to** prepare dinner. <TEC: Achievers A1+>

(22) Let's **try shouting** 'Lady Time' <TEC: New Green Line 1>

(23) He was never **afraid of making** a fool of himself, either. <TEC: Join the Team 3>

(24) On the cover, there is **a big woman wearing** goggles, a type of protection glasses, and a shield <TEC: Piece of Cake 3e>

(25) And another great thing about my job **is travelling**. <TEC: Green Line 2>

The remaining concordance lines were manually annotated for morphosyntactic features which could not easily be automated (see 4.2.2.4) as well as for the functional aspects detailed in 4.2.2.5.

4.2.2.4 Morphosyntactic features

Concordance lines were only annotated as questions if the progressive verb itself was the focus of the question, regardless of the syntax (26); the presence of a tag question following the progressive verb phrase was not coded as a question (27).

(26) **You're working tomorrow?** yes yes <BNC2014: S33B>

(27) you're not going swimming yet **are you?** <BNC2014: SQG4>

GET and HAVE passives were annotated as passives.

4.2.2.5 Functions of the progressive

Time reference occasionally proved more difficult to annotate than anticipated. For instance, the Youth Fiction corpus yielded many examples of free indirect speech, whereby the authors used time adjuncts associated with the present time reference to create a sense of immediacy in their narration of past events (28). Despite the contradicting adjuncts, such instances were nevertheless coded as 'past' time references.

(28) The smile **was getting** on his nerves, and **now** he was hearing things. <Youth Fiction: Pratchett 1999: The Fifth Elephant>

A further categorising decision had to be made concerning the interpretation of the time reference of progressives in reported thought and speech. Unfortunately, cases of tense shifting are not discussed in Römer's study (2005). It was decided that, in the present study, the variable time reference refers to the (presumed) intended time reference at the time the progressive verb phrase was first uttered or thought. Thus, excerpt (29) was coded for time reference 'present' and (30) for 'future'.

(29) I called out to them, saying **I was going** to bed. I didn't want them asking any awkward questions. <TEC: Access G 5>

(30) Finally, I asked if Benny had plans to visit more festivals. He told me that in May **he was attending** the Calaveras County Fair in California for the world's biggest frog jumping competition, with thousands of frogs. <TEC: Achievers B1>

Further, the core functions and additional functional features of the progressive that emerged from Römer's (2005: 80–110) corpus-driven investigation of the progressive in spoken British English were also coded for manually. Over the course of the first annotation cycle, a number of changes to Römer's (2005) categorisation were deemed necessary. First, the category 'general validity', which in Römer's (2005) annotation scheme was treated as an 'additional function' category, was transferred to the 'time reference' variable because it proved very difficult to logically assign a verb referring to a generally valid state or action to past, present or future time reference, e.g.:

(31) I am Plymouth's official town crier.
Ah, you call out the news so that everybody in the city knows **what's happening**.
That's correct. <TEC: Access G 2>

Second, due to low inter-rater agreement disambiguating between the two categories 'emphasis or attitude' and 'shock or disbelief', these were merged into one broader 'emphasis/shock' category. Finally, the additional function category 'old and new habits' was deleted because it frequently overlapped with the 'repeatedness' function and the two were thus difficult to disambiguate. Table 9 lists the additional function categories used for the present study.

Table 9: Core and additional function categories for the progressive (all examples from Römer 2006b: Appendix)

Function	Example sentence
Continuousness	the strange black things that some of **you are holding** in your hands are called riders, and these are end leaves for the storage binder
Non-continuousness	I say real progress has been made but today **I am asking** you to think about the next step
Repeatedness	she doesn't eat that much, but what **she is eating** i everything's sweet. Mm. Chocolates and Is she eating them? Pardon?
Non-repeatedness	Oh I see. Pork's very nice. **What's Geoff eating?** Sausage roll. Oh Geoff, you've only just had your tea!

Politeness or softening	I'm sorry I'm not clear as to whether **you are suggesting** that there should be policy upper case criteria and some non policy lower case criteria.
Emphasis/shock	**You're not suggesting** pregnancy's a disease there are you?
Gradual change and development	Er, it is a very difficult climate, **it's becoming** increasingly difficult, and indeed, it's affecting the work that we do
Framing	So anyway yesterday afternoon **I was checking** through it when the phone went again.

4.2.2.6 Validation of the annotation scheme

Having finalised the coding scheme, the present author and a student research assistant[22] each annotated a set of 100 random concordance lines from each of the four corpora under study. Inter-rater agreement was found to be 96.22% (across the nine manually coded variables). The majority of disagreements concerned the 'continuous/non-continuous' variable, which is particularly sensitive to subjective interpretation. Following discussions, it was agreed to follow Römer's (2005: 87–88) criteria ("procedural", "extension over a certain time span" vs. "short-term actions", "punctiform") and examples as closely as possible. As a result, (32) was coded as 'continuous' whilst (33)–(34) were not. Half the concordance lines were then annotated by the present author and the other half by the research assistant. Difficult cases were discussed to reach a common agreement and the label 'unclear' was applied for cases which lacked the necessary context to make a reliable decision.

(32) He didn't know what **was happening**, he couldn't explain anything. <TEC: Green Line 4>

(33) I'm not saying that **I'm** never **coming back**. <TEC: New Missions 2[de]>

(34) it's **you're putting** a T sound in there <BNC2014: SMHY>

Intra-rater agreement rates over a period of four weeks were 98.66% and 96.22% and, aside from a few inattention errors in either the first or second annotation phase, the differences were mostly due to the second coder seemingly gaining in confidence with fewer concordance lines being coded as 'unclear' in the 'continuous/non-continuous' and 'repeated/non-repeated' variables in the later sample.

4.2.2.7 Frequency of the progressive

This analysis aims to compare the frequency of progressives in the two TEC registers with that of their corresponding reference corpora. To this end, the simplest method is undoubtedly a comparison of relative frequencies per number of words or tokens (see (35), termed 'M-coefficient' in Smitterberg 2005). However, a meaningful comparison of the frequencies of progressives is difficult to achieve using such a

[22] Many thanks to Tatjana Winter for her meticulous work and insightful comments.

coefficient because it necessarily confounds the frequency of progressives with that of finite verb phrases.

(35) $M = N_{PROG} / N_{WORD} \times 100{,}000$

Preliminary analyses showed Textbook Conversation to be considerably more nominal than naturally occurring conversation (Le Foll 2017) so that M-coefficients would likely not suffice to conclusively answer the first research question (see 4.1.3). Moreover, attempting to conduct statistical tests to compare M-coefficients would make little sense because, in practice, language users do not choose between a progressive form and any other word. Instead, they are more likely to choose between a progressive or a non-progressive verb form (see also Mair & Hundt 1995: 114f). Furthermore, the test typically used for such comparisons, the chi-square test, requires the categories to be compared to be mutually exclusive. This is difficult to achieve in the case of progressives because a progressive verb phrase can span more than two words (e.g., *have been eating, are being wasted*) and thus the word/progressive categories are not necessarily mutually exclusive (see Smitterberg 2005: 39-42).

The K-coefficient (36) (devised by Nickel [1996] and discussed in Smitterberg 2005: 42-45 in detail) addresses these drawbacks by dividing the number of progressives by the number of verb phrases, excluding those that the progressive construction cannot fill.

(36) $K = N_{PROG} / (N_{VERB} - N_{NOPROG}) \times 10{,}000$

Whilst this approach satisfies the statistical requirement for a progressive vs. non-progressive alternation ratio measure, it is both difficult to operationalise without considerable time-consuming manual annotation and highly subjective. Indeed, decisions as to whether a progressive would theoretically have been possible in any given verb phrase are inherently theory-dependent. Consequently, such a measure would yield results which would not be comparable across studies. An arguably more objective measure is the V-coefficient (37) (Smitterberg 2005: 44–45) which relates the number of progressives to the number of verb phrases:

(37) $V = N_{PROG} / N_{VERB} \times 10{,}000$

Whilst the V-coefficient is fairly easy to approximate using a POS-tagged corpus, it retains a number of the disadvantages encountered with the M-coefficient. If non-finite verb phrases are not included in progressive count (as in this study and several others, e.g., Rautionaho 2014; Rautionaho & Deshors 2018; Rautionaho, Deshors & Meriläinen 2018), these also need to be excluded from the verb phrase count. Furthermore, if a variationist stance on the progressive/non-progressive construction is to be adopted, the number of progressives presumably ought to be subtracted from the total number of finite verb phrases. Smitterberg (2005: 45–48) also advocates for the exclusion of certain constructions in which the progressive is never featured, i.e.,

in imperative and BE *going to* + infinitive constructions. Automatically excluding imperatives proved computationally difficult, but since BE *going to* + infinitive constructions were captured by the CQL query, and manually tagged for separately, they could easily be counted. Thus, this study uses the following F-coefficient (38) ('F' for finite verb phrases) as a comparison measure between textbook and naturally occurring data:

(38) $F = N_{PROG} / (N_{FINITE\_VERB\_PHRASES} - N_{PROG} - N_{GOING\_TO}) \times 10{,}000$

To identify finite verb phrases (hereafter: FVPs), Rautionaho (2014: 78) manually annotated thousands of verb phrases. Though undoubtedly the more accurate method, in the context of this case study, relying on relatively large corpora, it was not a feasible option. However, the availability of fast and fairly robust automatic dependency parsers for English means that it is nowadays possible to automatically retrieve verb phrases on the basis of syntactic dependency relations. Consequently, a combination of POS and syntactic dependencies information was used to approximate the total number of FVPs in each of the corpora under study. To this end, the corpora were tokenised, parsed and tagged using the Python library {spaCy} (Honnibal & Johnson 2015; Honnibal & Montani 2017) via the R interface to Python {reticulate} (Allaire, Ushey & Tang 2019) and the R package {cleanNLP} (Arnold 2017).

First, POS tags were used to identify all verb forms. Second, the syntactic dependency parsing information capturing the relation connecting child to head was used to determine which verbs were the main verbs in FVPs (thus chiefly removing auxiliaries) and to exclude all verb forms from non-finite verb phrases. The R command which was employed to this effect is the following:

(39)
```
nrow(subset(spacy_parsed_tagged_corpus, upos=='VERB' & relation %in%
c('ROOT', 'relcl', 'ccomp', 'advcl', 'conj')))
```

Using this command, the following verb forms would first be retrieved from the paragraph above on the basis of their POS tag: *were, used, identify, capturing, connecting, was, used, determine, were, removing, exclude, was, employed* and *is.* However, only the following would then be counted as a result of their syntactic dependency relation (given in brackets): *used* (none = ROOT), *used* (none = ROOT), *were* (ccomp), *removing* (advcl), *exclude* (conj), *employed* (relcl) and *is* (none = ROOT). Hence, according to this approximation, the above paragraph is said to contain seven FVPs. As the example above demonstrates, this method is far from perfect, in particular due to the non-negligible error rate in the POS-tagging and dependency-parsing processes and the fact that the relations 'advcl' and 'conj' can cover both tensed and non-tensed verb phrases. Nonetheless, it was found to be the most reliable automatic method to approximate the number of FVPs for the purposes of this study.

4.2.3 Collostructional analysis

Collostructional analysis was designed to empirically investigate the interactions between grammatical constructions and the lexemes associated with them (Stefanowitsch & Gries 2003; Gries & Stefanowitsch 2004a; Gries 2015a; 2019a). Gries (2019a: 386) describes the method as an "extension of the notion of collocation", to be understood in the traditional sense of co-occurrences of two lexical items, but also of words and patterns (see, e.g., Hunston & Francis 2000) and of words and constructions (see, e.g., Goldberg 1995; 2006).

Collostructional analysis takes a construction as a starting point and measures which lexemes are attracted to or repelled by the construction by measuring whether they occur more or less frequently than expected on the basis of the total lexeme count in the corpus under study. The lexemes associated with a particular construction are referred to as the construction's collexemes (Stefanowitsch & Gries 2003: 215). By contrast, linear collocational approaches are usually based on the interpretation of patterns emerging from either the manual inspection of (semi)-automatically extracted concordance lines, or from lists of the node word's most frequent collocates, traditionally operationalised as words or lemmas most frequently found within a user-specified span around the collocation's node word. In such an approach, the semantic relation between the nodes and their collocates are not usually verified (but see, e.g., Uhrig & Proisl 2012 for a method advocating the use of dependency parsing to identify collocations). Collostructional approaches, on the other hand, restrict the semantic exploration to the syntactic frame(s) of the construction(s) under investigation, thus considerably reducing noise. Crucially, they also account for the marginal frequencies of each collexemes within the corpus under study (for demonstrations of the superiority of collostructional methods as compared to raw and basic relative frequency-based approaches, see Gries, Hampe & Schönefeld 2005; 2010; Desagulier 2014: 155–156; Gries 2015a).

Originally, three types of collostructional analyses were proposed (see Gries 2015a for an overview of the extended family of collostructional methods):

a) Simple Collexeme Analysis (Stefanowitsch & Gries 2003) is used to quantify the extent to which individual words/lemmas occurring in a construction are attracted to or repelled by that construction.
b) Co-varying Collexeme Analysis (CovCA) (Gries & Stefanowitsch 2004b; 2004a) attempts to quantify how much words/lemmas in one slot of a construction are attracted to or repelled by words in a second slot of the same construction (e.g., the into-causative: *to trick/force* [verb slot 1] *someone into buying/accepting* [verb slot 2] *something*) (examples from Gries 2019a: 386).
c) Distinctive Collexeme Analysis (Gries & Stefanowitsch 2004a) compares the extent to which words are attracted to two functionally similar constructions.

A fourth type, Key Collostructional Analysis (Gilquin 2012; 2015a; 2015b; 2016a) involves computing a Distinctive Collexeme Analysis, but rather than comparing the association of words/lemmas with two constructions, it is used to compare the association of words/lemmas with one construction across two or more corpora representing different language varieties (Gilquin 2016: 239–250). Gilquin applies this method to compare the phraseological patterns of use of causative constructions with MAKE (2012; 2016) and phrasal verbs (2015b) in native, EFL and ESL learner varieties. Without naming it as such, Deshors (2017) applies a similar method in her study of the progressive in Word Englishes. She extends it by comparing series of CovCAs to assess the association strength not only between individual lemmas and English varieties, but also between semantic domains and varieties.

Unlike simple frequency-based approaches, which generally report high-frequency light verbs as the strongest collocates, collostructional methods often reveal associations with semantically richer lexical verbs (see also Deshors 2017: 265) and can thus provide insights into semantic, functional and pragmatic domains that are attracted to or repelled by specific constructions.

In theory, collostructional analysis may be conducted with any of the many association measures frequently used in collocational research (see Manning & Schütze 1999: Ch. 5; Evert 2005: Ch. 3 for overviews); however, as pointed out in the aforementioned literature, many of these statistics assume normal distributions and homogeneity of variance, which have been shown to be unrealistic assumptions for natural language data (see Evert 2005: 2.3.1). Moreover, most of these measures perform particularly poorly, i.e., either vastly over- or under-estimating the strength of association, when dealing with very infrequent word co-occurrences. This is potentially highly problematic for collostructional analysis because, since construction type/token frequencies have been demonstrated to follow Zipfian distributions (see 1.5), we expect the vast majority of collexemes associated with any construction to occur at very low frequencies (Stefanowitsch & Gries 2003: 218).

In this thesis, collostructional analyses are conducted using Gries's (2019b) {coll.analysis} package for R with the default association measure setting of $-\log_{10}$ Fisher-Yates exact [FYE], one-tailed (except for the comparative DCAs, see 4.2.3.2). FYE test is often hailed as the significant test of choice for co-occurrence contingency tables (Pedersen 1996), in particular when low frequencies and skewed distributions are expected (Evert 2009: 1235–1236). Indeed, FYE does not rely on approximations which may skew results for low-frequency data, nor does it make specific sample size demands. This notwithstanding, FYE is known to be computationally expensive (Evert 2009: 1235); however, the {coll.analysis} version used here (v. 3.5) fixes the issue of infinite p_{FYE} collostructional strength (CL) values being returned for very strong associations or repulsions (cf. Schmid & Küchenhoff 2013: 537) by extending

the numerical computing abilities of the system running the R script thanks to the {Rmpfr} library (Maechler 2019).
In this chapter, two different collostructional methods are applied. The following subsection (4.2.3.1) describes how, inspired by Gilquin's method (2012; 2015a; 2015b; 2016a), two types of Co-varying Collexeme Analyses are used to compare, first the verbs used in the progressive, and second, their semantic domains (as in Deshors 2017) across the two TEC subcorpora and their corresponding reference corpora. Subsequently, 4.2.3.2 explains how the method is expanded to be able to combine the advantages of a Key Collostructional Analysis with those of Distinctive Collostructional Analyses.

4.2.3.1 Co-varying Collexeme Analyses: *Verb/Semantic domain + Corpus*

The first set of Co-varying Collexeme Analyses (thereafter, CovCA) conducted as part of this study contrasts the verb lemmas associated with a single constructional variant, the progressive construction, in two subcorpora of the TEC (Conversation and Fiction) and random samples from the Spoken BNC2014 and the Youth Fiction corpus. The second set of CovCAs focuses on the semantic domains to which these lemmas belong and their associations with the progressive construction.

Whilst some insights can only be gained with a progressive versus non-progressive alternation investigation (as in 4.2.3.2), the advantage of this single constructional variant is that phrasal and prepositional verbs can easily be investigated separately. For instance, it enables us to compare the strength of attraction to the progressive aspect for the following verbs: LOOK, LOOK AFTER, LOOK AROUND, LOOK FOR, LOOK FORWARD, LOOK OVER and LOOK UP. For practical reasons, all these verbs are agglomerated into the verb group LOOK in the context of the distinctive alternation collostructional analyses (DCA, see 4.2.3.2).

The CovCAs were also carried out using Gries's (2019b) {coll.analysis} 3.5 for R. The input file for the first Conversation CovCA consists of a table of 4,846 rows where each row corresponds to one annotated concordance line featuring a progressive as defined in 4.2.2.3. The table has two columns: one for the verb lemma and the other for the (sub)corpus from which the concordance was retrieved. From this input table, the {coll.analysis} script constructs, for each of the 623 unique lemmas that are found in the Conversation progressive dataset, a contingency table like the one displayed in Table 10.

Table 10: Example contingency table for the verb SAY in the first CovCA conducted (see 4.2.3.1)

	SAY in the progressive	*Other lemmas in the progressive*	*Row totals*
Textbook Conversation	29 (a)	2,394 (b)	2,423 (a + b)
Spoken BNC2014 sample	153 (c)	2,270 (d)	2,423 (c + d)
Column totals	182 (a + c)	4,664 (b + d)	4,846 (a + b + c + d)

For each verb, the programme computes a collostructional strength (CL) value which, in these CovCAs, is a measure of the degree of association/repulsion between a verb lemma in the progressive and a corpus (sample). It is calculated on the basis of a comparison between a verb lemma's expected frequency in the progressive in a corpus (sample) if the null-hypothesis were true, i.e., under the assumption that the variables are independent of each other.

4.2.3.2 Comparative Distinctive Collexeme Analysis

This section describes the methodology for investigating the lemmas occurring as the main verb in FVPs most strongly associated with either progressive or non-progressive constructions within each corpus. To this end, separate progressive vs. non-progressive DCAs are conducted for each corpus (sample) and the CL values are then used to compare the results across the textbook and reference (sub)corpora for each register. In other words, a pair of progressive/non-progressive DCAs is compared for each of the two registers, conversation and fiction.

A known methodological problem in conducting DCAs is the operationalisation of the non-occurrence of the construction under study (see Bybee 2010: 98; and responses in Gries 2012: 488; Gries 2015a: 511); in other words, in this case, operationalising what constitutes a non-progressive construction. Intuitively, the best way forward is to identify "a level of resolution on which to count constructions that is close to the phenomenon in question" (Gries 2015a: 511). At the most basic level, this could be the number of words, or better verbs. However, this same problem was essentially already encountered in 4.2.2.7, where a more refined solution involving the automatic extraction of finite verb phrases (FVPs) using POS-tagging and dependency parsing information was presented. Thus, this method will also be used in the comparative DCAs, such that FVPs will serve as the count unit for all progressive and non-progressive constructions in an alternation paradigm.

Crucially, in order for the CL values of the two parallel DCAs of any one register to be comparable, it is necessary to ensure that the number of FVPs in the two corpora be equal. This is because the log-likelihood test which forms the basis of the CL calculation is highly sensitive to sample size. As a result, the data collecting and pre-

processing procedure for the DCAs is more complex than for the CovCAs. The following section describes the procedure developed to arrive at approximately the same number of FVPs per corpus (sample) for each pair of DCAs.

The total FVP verb lemma counts were automatically computed across the entire corpora using POS and dependency information inferred by the Spacy model (following the same procedure as in the calculation of the F-coefficients of progressives, see 4.2.2.7). The counts of the progressive FVPs, on the other hand, were the result of the manual annotation procedure (see 4.2.2.3), which, in the case of the reference corpora (i.e., in the Spoken BNC2014 and the Youth Fiction corpus), was only carried out for a random subset of all the progressive CQL query results. Consequently, it was necessary to calculate suitable ratios by which to divide the total lemma frequency counts across the full reference corpora in an attempt to match the sample size of their respective TEC subcorpus. Evidently, taking a simple word or token ratio bears the risk of strongly distorting the CL values – not least because the Spoken BNC2014 features proportionally more verbs than Textbook Conversation. However, if we assume that FVPs are evenly distributed across the reference corpora (which visual inspection the corpus dispersion plots from the Sketch Engine interface suggests is a valid assumption), we may divide the total number of FVPs in Textbook Conversation (64,292) by the number of FVPs in the Spoken BNC2014 (1,751,040) to obtain an operational sample size ratio (0.0367). This ratio can then be used to calculate the number of random concordance lines retrieved using the CQL query which need to be manually sorted and annotated and ultimately included in the DCA in order to match the number of FVPs from Textbook Conversation. In other words, out of the 126,395 hits returned by the progressive CQL query for the entire Spoken BNC2014, a random sample of 4,641 (= 126,395*0.0367) was manually sorted and annotated. Post-manual filtering (see 4.2.2.3), this random sample yielded 3,444 progressives, which is the figure listed in the corresponding cell of the example contingency table for the Spoken BNC2014 sample (see Table 13).

Consequently, it was possible to draw up lists of all the lemmas found in the manually annotated progressive concordance lines from both a subcorpus of the TEC and its corresponding reference corpus. With the computational tools described above, however, automatically identifying FVPs of phrasal and prepositional verbs with their particles was not technically feasible. Consequently, unlike in the CovCAs described in 4.2.3.1, the DCAs do not differentiate between verbs such as LOOK AFTER, LOOK UP and LOOK FORWARD TO. Thus, the first data preparation step consisted in reducing the verb lemmas in the annotated progressive concordance lines to their main verbs, i.e., dropping the prepositions and particles of multi-word verbs. Two lists of these single-word verb types, one for each register under study, were then compiled.

Subsequently, each list was used to count the number of FVPs with these lemmas as their main verbs in the corresponding TEC subcorpus and its full reference corpora. Since the FVPs had already been automatically extracted previously (see 4.2.2.7), this could easily be realised with a for-loop script. The resulting count tables consist of a row per verb lemma from that register's lemma list, a column for the number of FVPs in the TEC subcorpus of that register, and a column for the number of FVPs in the corresponding reference corpus (see
Table 11 for an extract of such a table).

Table 11: Last three entries of the count table for the conversation register before normalisation

Lemma	FVPs in Textbook Conversation	FVPs in the Spoken BNC2014
WORRY	142	1,043
WRITE	182	2,840
YAWN	0	15

Next, this dataset with the total number of FVPs per lemma in each corpus was combined with the count lemma information gathered in the annotated progressive concordance tables (see 4.2.2). For the TEC subcorpora, this information theoretically suffices to compute individual contingency tables for each verb, with counts for both progressive and non-progressive verb lemmas as main verbs in FVPs, from which the CL of each lemma can be calculated. Table 12 presents an example contingency table for SAY in the Textbook Conversation subcorpus. The values in bold have been directly counted in the corpora. The others are inferred.

Table 12: Contingency table for SAY in Textbook Conversation (the figures in bold were directly counted from the subcorpus)

	SAY as the main verb of an FVP	*FVPs with verbs other than SAY*	*Row totals*
Progressive	**29**	2,394	**2,423**
Non-progressive	788	53,796	54,584
Column totals	**817**	56,190	**57,007**

In the case of the TEC subcorpora, all progressives FVPs were manually filtered from the complete progressive CQL query results (see 4.2.1), but the FVPs were extracted and analysed for their main verb lemma automatically. These different methods led to four lemmas (LOOP, NOURISH, OVERCHANGE and UPSET) that had been manually annotated as main verbs of FVPs in the progressive concordance lines, not being subsequently identified as verbs by the Spacy model and thus not counted as lemmas in FVPs. Therefore, when subtracting the number of progressives from the total number of FVPs for these four lemmas, the non-progressive count returned a negative value. Since this is clearly non-sensical, these counts were adjusted to one, to reflect the fact that one progressive occurrence of each was in fact observed in this subcorpus.

This adjustment resulted in an increase of the total number of FVPs in Textbook Conversation to 57,007, as displayed in Table 12.

Note that, contrary to the grand total number of FVPs used to calculate the ratios described above, the FVPs row + column totals listed in the contingency Tables 12 and 13 reflect the number of FVPs with lemmas from the lemma lists compiled for each register (which include all the lemma types found in the progressive concordance lines of a textbook subcorpus and its corresponding reference corpus sample). This explains why the total of 64,292 FVPs in Textbook Conversation was first used to calculate a sampling ratio, but we now only see a total of 57,007 FVPs in Table 12. In the following, this total is referred to as the total number of 'list-FVPs'.

As mentioned at the beginning of this section, in order to be able to compare two DCAs, the grand totals of all the contingency tables compared ought to be the equal. To achieve this, the number of list-FVPs from each textbook subcorpus was divided by the number of list-FVPs from its corresponding reference corpus sample. For conversation, this resulted in the following ratio: 57,007/1,526,439 = 0.0373. In other words, using the Spacy model, a total of 50,045 FVPs with SAY as a main verb were counted across the entire Spoken BNC2014, but applying the ratio above reduced this total to 1,869.001, rounded off to 1,869 in Table 13.

Table 13: Contingency table for SAY in the Spoken BNC2014 sample (the figures in bold have been directly counted, underlined approximated from the corpus)

	SAY as the main verb of a FVPs	*Total verb lemmas of FVPs except SAY*	*Row totals*
Progressive	**214**	3,230	3,444
Non-progressive	1,655	51,908	53,563
Column totals	1869	55,138	**57,007**

However, similarly to the TEC subcorpora, these contingency tables for the reference corpora also included a number of lemmas for which the total number of non-progressive forms was negative. In some cases, this is presumably for the same reason as for the TEC subcorpora, i.e., that the Spacy model did not recognise certain lemmas as verbs. Looking at the list of the 131 lemmas with negative non-progressive counts in the Spoken BNC2014, this would appear to be the case for a number of lemmas whose noun form is typically more frequent than its verb form, e.g., CONTACT, PARTY, PROTEST, QUEUE, RACE and SWEAT, and which were, as a result, presumably disproportionately rarely tagged as verbs by the Spacy model.

In the majority of cases, however, these negative values are largely due to low-frequency lemmas which were observed in the sample of the 3,444 progressive concordance lines from the Spoken BNC2014 and thus included in the conversation

lemma list but, because they are very low frequency, whose post-normalisation FVP counts (i.e., after having applied the list-FVP ratio described above) were below one. The problem is demonstrated with the lemma ACTION which was observed once in a progressive construction among the sample of 3,444 progressives from the Spoken BNC2014 in the context of telling an anecdote about a complaint at work (40).

(40) cos he should do that and then just sent he should have sent a message through to either Jill
yeah
or John the new guy
yeah
and said **I'm not actioning** these just to let you know <BNC2014: SVD6>

According to the formula applied to track FVPs using the Spacy model (see 4.2.2.7), the extract printed above (40) is the only occurrence of the verb lemma ACTION across the entire Spoken BNC2014. Thus, the total number of FVPs with this ACTION as its main verb lemma for the full corpus is equal to one but, as a result of sampling ratio, only 0.0373 for the sample. Consequently, a negative non-progressive count is returned when subtracting the number of progressives from the total number of FVPs (see Table 14 for further examples of such lemmas). It goes without saying that for some of these lemmas, a combination of the first and second causes may well have contributed to negative non-progressive scores (e.g., infrequent word forms which, when they do occur, are unlikely to be verbs and thus not recognised as such by the Spacy model, e.g., ASPHALT or REV, both from the Spoken BNC2014 list).

Table 14: Examples of lemmas whose non-progressive FVP counts were negative in the Spoken BNC2014

	Total FVPs (full corpus)	*FVPs (in sample)*	*Progressive FVPs (in sample)*	*Non-progressives FVPs (in sample)*
ACTION	*1*	*0.04*	*1*	*-0.96*
PARTY	*15*	*0.56*	*1*	*-0.44*
SHINE	*19*	*0.71*	*1*	*-0.29*
SWEAT	*92*	*3.44*	*5*	*-1.56*

In order to resolve this issue, removing all lemmas whose reference corpus non-progressive counts were negative and whose progressive counts were equal to just one (e.g., for the Spoken BNC2014, lemmas such as those displayed in Table 14) was considered. This makes sense because the raw frequencies are, in any case, too low to reach statistical significance when conducting a DCA. However, eliminating all of the lemmas corresponding to these criteria would risk overlooking lemmas which are proportionally over-represented in the corresponding TEC subcorpora. These include, for instance, the lemma SHINE, which was observed ten times in Textbook Conversation and which is therefore a potentially interesting candidate for the comparison of the DCAs in the conversation register. Consequently, only lemmas which fulfilled the criteria above and whose total FVP counts in the corresponding

TEC subcorpus was equal to zero were excluded from the count tables to be used for the DCAs. For the conversation DCAs, this resulted in the elimination of 97 unique verb lemmas (out of the original 589). For the few remaining lemmas whose FVP counts were still negative (e.g., SHINE), the number of FVPs was adjusted to match that of the number of progressives with that same lemma. Table 15 shows the data displayed in Table 14 once these adjustments had been made.

Table 15: Example lemma frequency counts for the Spoken BNC2014 after having adjusted for sampling/tagging errors

	Total FVPs (full corpus)	*FVPs (in sample)*	*Progressive FVPs (in sample)*	*Non-progressives FVPs (in sample)*
PARTY	*15*	*1*	*1*	*0*
SHINE	*19*	*1*	*1*	*0*
SWEAT	*92*	*5*	*5*	*0*

Following the data wrangling described above, the total number of list-FVPs included in the conversation DCAs corresponded to 57,006.63 (see Table 16).

Table 16: Contingency table for SAY in the Spoken BNC2014 post-adjustments

	SAY as the main verb of a FVPs	*FVPs with main verb other than SAY*	*Row totals*
Progressive	**214**	3,133	3,447
Non-progressive	1655.00	52,004.63	53,659.63
Column totals	1869.00	55,137.63	**57,006.63**

There remains a difference of 0.37 more FVPs in Table 12 (Textbook Conversation) than in Table 16 for the corresponding Spoken BNC2014 reference sample (a difference corresponding to 0.0006%). Fortunately, Gries's (2019b) {coll.analysis} script allows the user to manually enter row totals for the contingency tables that diverge from the actual totals in the datasets. It is therefore possible to adjust the row totals of the contingency tables fed into the formula so that, for the purposes of the CL values computation, the total number of FVPs are matched exactly. Thus, 57,007 was entered as the total number of FVPs for the Spoken BNC2014 analysis. The same procedure was followed for the comparison of the progressive vs. non-progressive DCAs computed from the Textbook Fiction and Youth Fiction data.

A final issue concerns the choice of association measure for the DCAs. Fisher's Exact test (FET) can only be performed with integer values. Therefore, all DCAs are conducted using the log-likelihood ratio (LLR, G^2) to calculate the collostructional strength (CL) values. The LLR is calculated as shown in (41), where expected values are computed from row and column totals of the contingency table (Gries 2019a: 387).

(41) $2\sum_{a}^{d} obs \quad log\frac{obs}{exp}$

LLR is currently the second most widely used association measure in collostructional analysis (Gries 2019a: 386). Like FET, it does not rely on or make any distributional assumptions, nor does it make specific sample size demands. In fact, the two measures are highly correlated and Evert (2009: 1235) demonstrated that "log-likelihood provides an excellent approximation to association scores computed by Fisher's test".

Controlling for the total number of FVPs allowed for a better comparison of the results of the parallel DCAs because the log-likelihood ratio (G^2) association measure which was chosen as the basis of the CL calculation is highly sensitive to sample size (like other association measures, see e.g., Gries 2019a: 388–390). The results of the DCAs were first analysed individually. Then, the CL values were standardised and visually compared with scatterplots.

All quantitative analyses were supplemented with extensive close reading of the relevant concordance lines, examples of which are given throughout the following results sections.

4.2.4 Correspondence Analysis

Following Deshors (2017), Correspondence Analysis (CA) was also used to explore the relationships between three of the dimensions under study, namely 'Semantic domain', 'Variety' (Textbook English vs. ENL) and 'Register' (Conversation vs. Fiction). CA is an extension of Principal Component Analysis and is used for exploratory analysis of interactions among categorical variables. It enables us to simplify, summarise and visualise complex contingency tables on two-dimensional plots (Desagulier 2017: 257–269; Husson, Lê & Pagès 2017: Ch. 2). Since the present study only deals with two different registers and two different varieties, a binary CA with the variables 'Semantic domain' and 'Corpus' suffices to effectively summarise the data (unlike in Deshors' [2017] study where a Multiple Correspondence Analysis was necessary).

As a bottom-up approach, CA makes no prior assumptions concerning the groupings to be unveiled in the data; however, the variables must be globally interdependent for the method to be valid. To this end, a chi-square test of independence of deviation between the two variables must be conducted before proceeding (Desagulier 2017: 261). The CA itself outputs factor scores for both the row points and column points of the contingency table (corresponding here to a 'Semantic domain' by 'Corpus' cross-tabulation). The scores computed then function as the coordinates for the biplot which helps to visualise potential associations between the variables, whereby "two rows or two columns have a similar profile if they are relatively close in the Euclidean space" (Desagulier 2017: 266).

The intensity of the relationship is measured with Cramér's V (Husson, Lê & Pagès 2017: 82) where a score of 0 would indicate no association, whereas as 1 would point to the (highly unlikely!) exclusive association of semantic domains of progressives with the variable 'Corpus'. In CA, Cramér's V reflects the degree of variance within the contingency table (Desagulier 2017: 263).

The correspondence analysis was conducted with the default settings of the CA function of the R package {FactoMineR} (Lê, Josse & Husson 2008).

4.3 Results and discussion

4.3.1 Frequency of progressives

For the Textbook Conversation subcorpus the relative frequency of progressives, as approximated using the F-coefficient method described in 4.2.2.7, is 395.48 progressives per 10,000 finite verb phrases. By contrast, it was found to be 557.98 per 10,000 finite verb phrases in the Spoken BNC2014 sample. Thus, even when accounting for the lower number of verbs and finite verb phrases in Textbook Conversation (see 4.2.2.7), the F-coefficient of progressives is significantly lower in Textbook Conversation than in the reference ENL conversation corpus ($\chi^2(1) = 191.81$, $p < 0.001$). Though the difference appears impressive, the strength of the association is in fact very small ($\varphi = 0.038$). Nonetheless, this finding contradicts the claim made by Biber & Reppen (2002b: 203) that the progressive aspect is overrepresented in textbook representations of natural conversation.

The Textbook Fiction subcorpus, on the other hand, has a slightly higher relative frequency of progressives than its corresponding reference corpus, the Youth Fiction corpus. However, this numerical difference is neither significant at the threshold of $p < 0.01$, nor meaningful in terms of its effect size ($\chi^2(1) = 5.7508$, $p = 0.015$, $\varphi = 0.009$).

4.3.2 Morphosyntactic features

This section focuses on morphosyntactic features of the progressive constructions found in Textbook Conversation and Fiction. The features have been chosen on the basis of previous studies on the progressive aspect (in particular Römer 2005; Rautionaho 2014; and Deshors 2017). Raw figures are comparable across the two varieties (Textbook vs. naturally occurring ENL) because the following analyses are based on all the progressives retrieved from the Textbook Conversation and Textbook Fiction texts, as compared to the same number of random progressive concordance lines from the Spoken BNC2014 and the Youth Fiction corpus, 2,423 and 1,517 respectively.

4.3.2.1 Tense, aspect and modality

This first section focuses on the distributions of tense forms of the progressive which, for the purposes of this analysis, also include the perfect aspect and modal forms. Römer (2005: 244) reported that EFL textbook dialogues vastly overuses the present progressive. The results displayed in Fig. 4 would, at first sight, appear to back this finding for modern school EFL textbooks since the present progressive accounts for 75.65% of all progressives in Textbook Conversation, as opposed to 62.65% in the Spoken BNC2014 sample. Furthermore, Römer (2005: 271) reports an under-representation of present perfect and past perfect progressives with a number of high-frequency verbs (e.g., ASK, GO, PLAY and STAY). Since the proportional usage of both past and present perfect progressives is very low, the two forms were merged into one 'perfect' category. The present results (see Fig. 4) suggest that there is, indeed, an under-representation of the past progressive in Textbook Conversation. By contrast, perfect progressives seem, if anything, to be over-represented compared to the Spoken BNC2014 sample.

The final category, entitled 'modal', contains progressives combined with CAN, COULD, HAVE TO, MAY, MUST, SHOULD, WILL and WOULD. Though still very marginal, the proportion of modal progressives relative to other progressive forms is almost twice as high in the Spoken BNC2014 sample (3.76 %) as in Textbook Conversation (2.15 %). We might have expected modal progressives in Textbook Conversation to feature a much more restrained range of modals; however, the full range is covered, thus with very low frequencies of occurrence for each modal per textbook series. In contrast, WILL and WOULD dominate in the Spoken BNC2014.

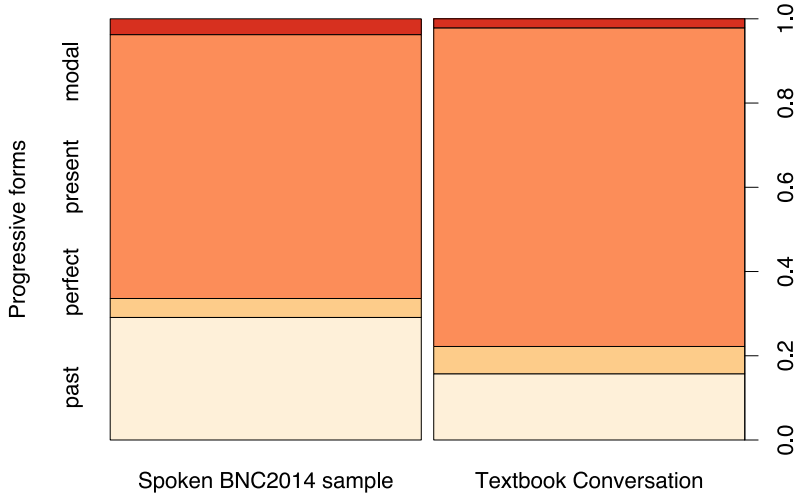


Fig. 4: Distribution of progressive forms in Textbook Conversation and the Spoken BNC2014

The distributions are significantly different ($\chi^2(3) = 145.57$, $p < 0.001$)[23] and whilst the strength of the association is small, it nevertheless suggests a meaningful association worth considering (Cramér's $V = 0.173$). Fig. 5 is an association plot (Meyer et al. 2003) which displays Pearson residuals for each tense form category as bars (a positive residual corresponds to an observed frequency being greater than expected). The intensity of the shading corresponds to its relative importance whereas the width of the bars represents the squared root of the expected value of each category. Thus, contrary to what Fig. 4 may have led us to believe, Fig. 5 shows that the most striking difference in the distribution of tense forms is in fact in the past progressive. Standardised Pearson residuals (as recommended by Agresti 2002: 81 and cited in Levshina 2015: 220-221) confirm that, whilst all categories make significant contributions to the χ^2 statistic at the significance level of 0.01, with a standardised Pearson residual of 11.16, past progressive is the category with by far the greatest contribution.

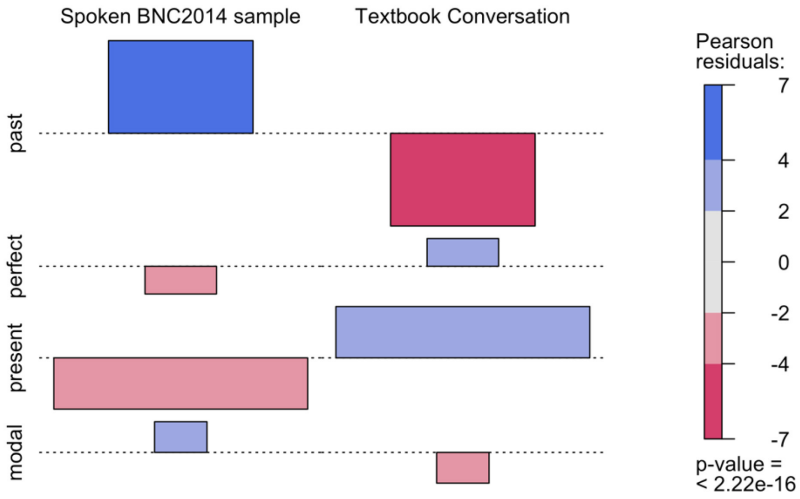


Fig. 5: Association plot comparing the progressive form distributions in Textbook Conversation and the Spoken BNC2014 sample

However, it seems problematic to examine tense form distributions across the entire Textbook Conversation subcorpus when we know that textbook authors are unlikely to introduce more complex forms like past and perfect progressives until a later stage. Consequently, Fig. 6 displays the tense distributions subdivided by textbook level (level A corresponding to the first year of EFL learning in lower secondary school, see 3.3.1). As expected, the beginner textbooks almost exclusively feature present progressive forms whilst the perfect and modal progressives are not introduced until

[23] All χ^2 tests conducted for this chapter were performed using the default settings of the chisq.test function in R. As such, Yates' continuity correction was applied for all 2 by 2 tables. A prerequisite for χ^2 tests is that the observations are independent. Although now a standard test in qualitative corpus linguistics, it could be argued that this condition is not met since many concordance lines come from the same textbook (series) and some from the same text. The same is true of the reference corpora, though this issue is much more limited due to the random sampling of concordance lines.

the third year of EFL instruction. Once they have been introduced, perfect progressives appear to be over-represented in Textbook Conversation. Modal progressives, on the other hand, are introduced more gradually. Whilst Fig. 6 shows that the present progressive is, in fact, not over-represented in Textbook Conversation, it is clear that past progressives, on the other hand, are under-represented across all textbook levels.

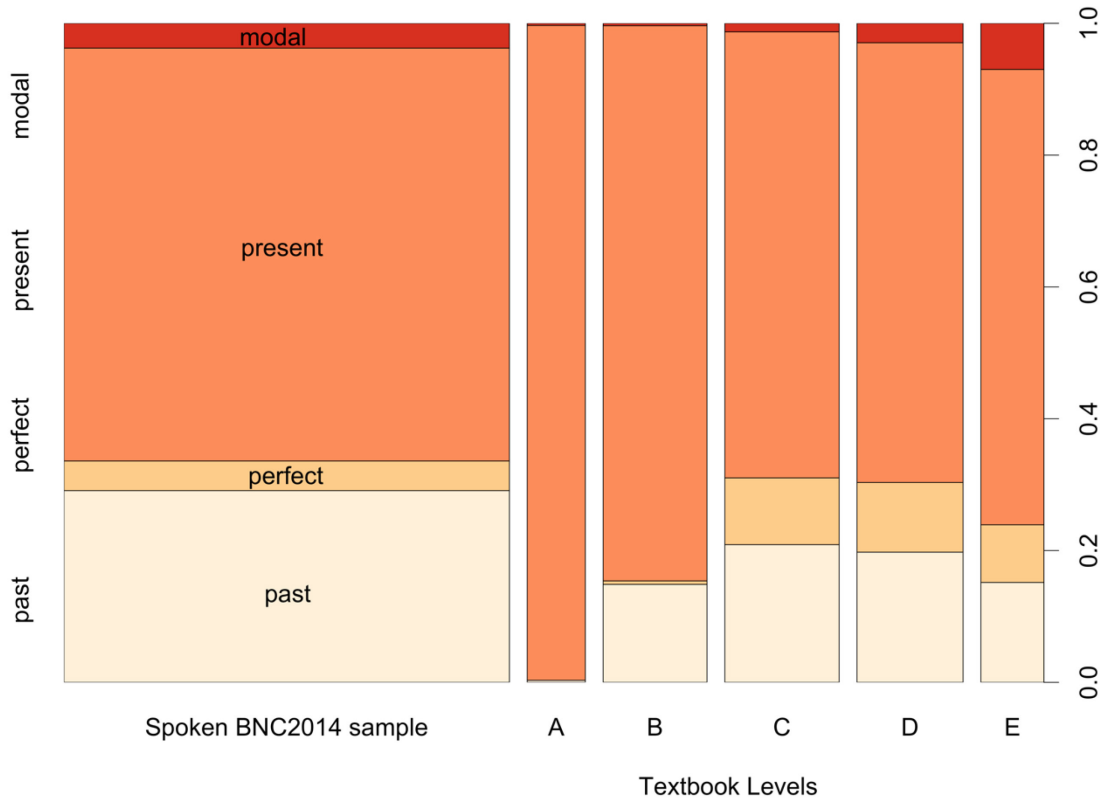


Fig. 6: Progressive form distributions in the Spoken BNC2014 sample and the five different Textbook levels[24]

The same analyses as above were replicated for the Textbook Fiction and Youth Fiction progressive concordance lines. As shown in Fig. 7, fewer discrepancies in tense form distributions were found. The strength of association between the tense distribution and the corpora is even weaker than in the previous analysis (Cramér's $V = 0.132$). As in Textbook Conversation, the fact that the present progressive is seemingly over-represented in Textbook Conversation ensues from the late introduction of other progressive tense forms in the textbook series. Indeed, the distribution of progressive tense forms in level D and E textbooks corresponds well to that observed in the Youth Fiction. Modal progressives are not introduced until level D and, in the most advanced textbook of each series, the proportion of modal progressives in Textbook Fiction texts matches that of the reference fiction corpus.

[24] Note that the level E textbooks do not feature fewer progressives in their dialogues. The bar representing level E textbooks is narrower because the Textbook Corpus contains two series which only go as far as level D (*Solutions* and *Piece of Cake*).

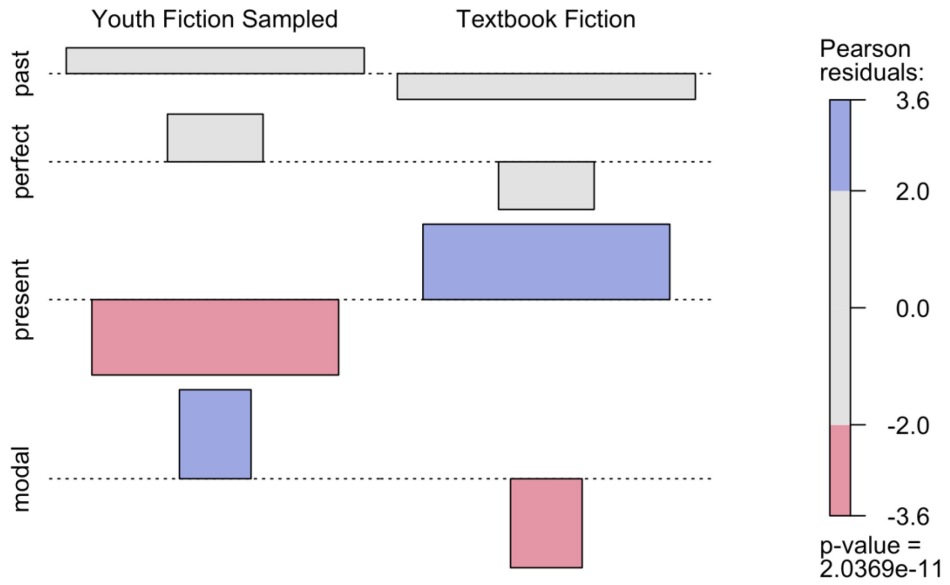


Fig. 7: Association plot comparing the tense form distributions in Textbook Fiction and the Youth Fiction sample

4.3.2.2 Contractions

In naturally occurring spoken English, verb contractions are very frequent and, indeed, out of the 2,423 progressive concordance lines extracted at random from the Spoken BNC2014, over half (n = 1,285) feature a contracted form of the auxiliary BE. This rate has seemingly not changed over the past few decades: Römer (2005: 245) reported very similar proportions for her sample of progressives from the Spoken BNC1994 and the spoken section of the Bank of English.

Taken as a whole, the proportion of contracted forms in the Textbook Conversation subcorpus is only marginally lower (46.43%) than that of the reference corpus (53.03%). Though the difference is significant ($\chi^2(1) = 20.868$, $p < 0.001$), the strength of the association between the proportion of contracted forms and language variety (Textbook English vs. ENL) is negligible ($\varphi = 0.066$). However, a differentiated view of the proportions per Textbook series reveals a different picture (see Table 17). In the present study, the two lowest proportions of contracted forms come from French publications: *Join the Team* (29.41%) and *Piece of Cake* (39.26%).

Table 17: Proportion (and raw numbers) of contracted forms of the auxiliary BE in progressive constructions

	Contracted forms	Full forms
Spoken BNC2014 sample	53.03% (1285)	49.97% (1138)
Access	54.07% (226)	45.93% (192)
Achievers	42.57% (43)	57.43% (58)
English in mind	57.02% (138)	42.98% (104)
Green line	43.17% (60)	56.83% (79)
Hi there	44.13% (94)	55.87% (119)
Join the team	29.41% (35)	70.59% (84)
New green line	44.50% (182)	55.50% (227)
Piece of cake	39.26% (64)	60.74% (99)
Solutions	45.72% (283)	54.28% (336)

Fig. 8 displays the proportions of contracted to full forms per progressive form. It shows that, in both Textbook Conversation and the Spoken BNC2014, contracted forms occur considerably more frequently in present progressives than in any other progressive form. More than three quarters of all present progressives in the BNC sample are contracted forms. This proportion is considerably lower in Textbook Conversation. It is also notable that, whilst perfect progressives are proportionally more frequent in Textbook Conversation, the vast majority are in non-contracted forms, whereas around half of naturally occurring perfect progressives feature a contraction. Past progressives, on the other hand, are very rarely contracted and this is well reflected in the textbook data.

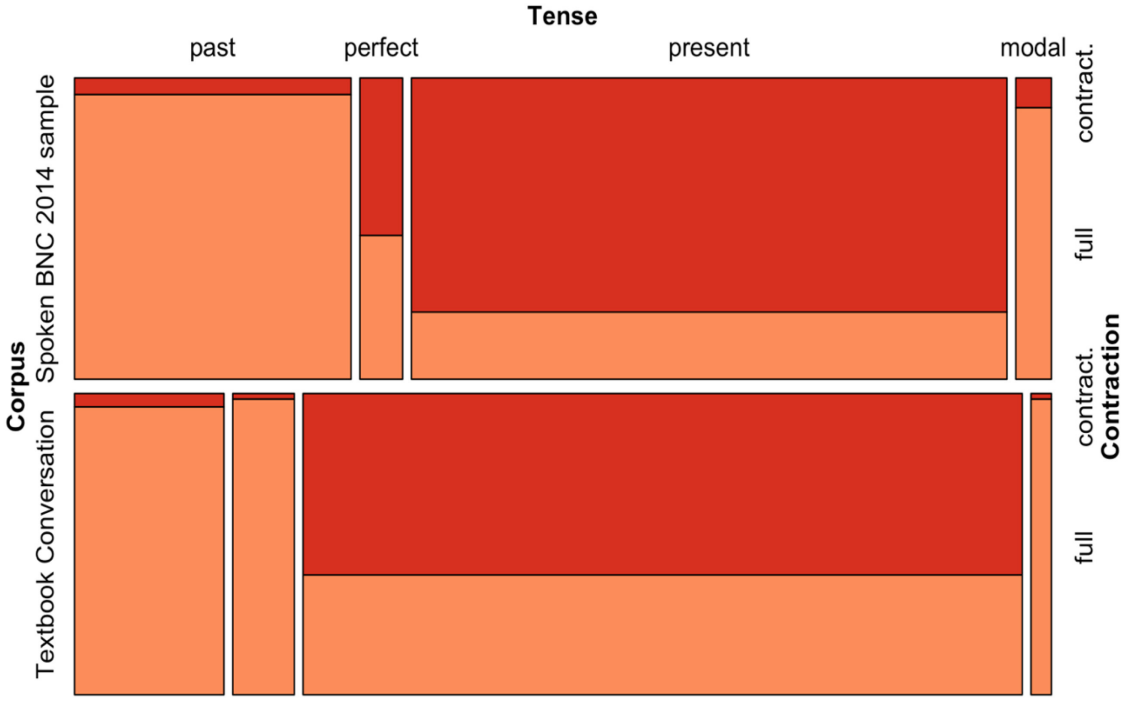


Fig. 8: Proportion of contracted progressives by progressive form in Textbook Conversation as compared to the Spoken BNC2014 sample

Compared to Conversation, narrative writing features considerably fewer contracted progressive forms. No significant difference in the proportion of contracted progressives between the Textbook Fiction subcorpus and the Youth Fiction reference corpus ($\chi^2(1) = 0.097478$, $p = 0.75$) was observed (see Table 18).

Table 18: Proportions of contracted and non-contracted progressives in Textbook Fiction and the Youth Fiction corpus

	Contracted form	Full form
Youth Fiction sampled	21.23% (322)	78.77% (1195)
Textbook Fiction	20.70% (314)	79.30% (1203)

As discussed in 3.3.1.4, a few textbook series hardly feature any narrative texts (e.g., *Piece of Cake* and *Solutions*). Disregarding those, the series with the lowest proportion of contracted progressives in Textbook Fiction remains, as in Textbook Conversation, the French series *Join the Team* (13.33%). *Achievers* also has a low rate (15.26%). The highest rates are found in *Hi There* (32.73%) and *Access* (26.76%).

The distributions of contracted auxiliaries across tense forms and subject person forms (*I*, *you*, *he/she/it*, etc.) in the two TEC subcorpora are very similar to that of the reference corpora.

4.3.2.3 Negation

Though the proportion of negated progressives observed in Textbook Conversation is marginally lower than in the Spoken BNC2014 sample (see Table 19), no meaningful quantitative difference could be established within the present datasets ($\chi^2(1) = 4.1278$, $p < 0.001$, $\varphi = 0.03$). This was also true of the Textbook Fiction vs. Youth Fiction comparison (see Table 20; $\chi^2(1) = 8.8826$, $p = 0.003$, $\varphi = 0.06$). Nor were any significant differences between the different Textbook series found.

Table 19: Proportion of negated progressives in Textbook Conversation and the Spoken BNC2014 sample

	Negated forms	Non-negated forms
Textbook Conversation	6.48% (157)	93.52% (2266)
Spoken BNC2014 sample	5.08% (123)	94.92% (2300)

Table 20: Proportion of negated progressives in Textbook Fiction texts and the Youth Fiction sample

	Negated forms	Non-negated forms
Textbook Fiction	3.10% (47)	96.90% (1470)
Youth Fiction sample	5.34% (81)	94.66% (1436)

However, qualitatively, it would appear that negation occurs in quite different functional and semantic contexts. Thus, in Textbook Conversation, negated progressives are typically found where an action is not taking place (42) or a state is negated (43). By contrast, natural conversations feature more misunderstandings which call for clarifications, and these frequently rely on negated verbs, e.g., (44)–(46).

(42) What's he doing? I don't know he **isn't answering** his phone. <TEC: Solutions Pre-intermediate>

(43) Sue, you **aren't wearing** your sports kit. P.E. starts in ten minutes. <TEC: Solutions Intermediate>

(44) I **am not talking** about finding a job <BNC2014: S2FQ>

(45) hold on we**'re not using** that one <BNC2014: S632>

(46) it's not crazy logic and I**'m not being** funny in the odd way <BNC2014: SG2Y>

Moreover, the topics dealt with in school EFL textbooks tend to avoid controversy and the vast majority of dialogues thus depict positive situations and conversations between people who generally agree with each other. This observation echoes the conclusions of earlier textbook studies (e.g., Rinvolucri 1999; Wajnryb 1996). Indeed, to Wajnryb (1996, cited in Timmis 2016: 5), the world depicted in textbooks is "safe, clean, harmonious, benevolent, undisturbed, and PG-rated". By contrast, perfectly harmonious interactions are clearly not always to be found in authentic conversational data and this is likely another factor that contributes to substantial qualitative differences in the type of negation found in the Spoken BNC2014 (47)–(49) as opposed to Textbook Conversation.

(47) yeah yes we**'re not looking** forward to that <BNC2014: SV4W>

(48) I'**m not trying** to be very mean but she sort [sic] of people that just wanna do it by themselves like their own way and don't listen <BNC2014: S7WY>

(49) you **are not buying** dodgy tablets off ebay or anything? <BNC2014: SPZR>

4.3.2.4 Questions

Römer (2005: 252) observed that almost twice as many progressives in interrogative contexts are found in textbook dialogues than in naturally occurring speech (20.48% versus 10.77%). This trend is confirmed in the present data set, with 17.91% of all progressives in the Textbook Conversation subcorpus being in the form of questions, compared to 11.39% in the Spoken BNC2014 sample. The difference in proportion is statistically significant ($\chi^2(1) = 40.379$, $p < 0.001$), but the strength of the association is very small ($\varphi = 0.092$). Once again, the most notable differences are qualitative in nature. Textbooks feature many more closed questions in the progressive (50) and

questions usually query what is happening right now, even in cases where the context of a shared environment probably ought to render this unnecessary (51).

(50) **Are you swimming** in the sea? **Are you playing** golf in the jungle? <TEC: Achievers Pre-intermediate>

(51) Jenny: What **are you doing**?
Terry: I'm watching TV.
Jenny: Why **are you watching** TV? You usually play football on Saturdays.
Terry: But it's raining. <TEC: Piece of cake 6e>

By contrast, questions in the progressive in the Spoken BNC2014 are frequently rhetorical (52) or serve to ask for confirmation (53)–(1) and often do not follow the "standard" question syntax typically taught in school textbooks, e.g., (54) and (1).

(52) okay get it recorded I**'m guessing**? yeah excellent no one is actually listening to this picking out words okay? <BNC2014: SG87>

(53) **isn't she having** another baby? <BNC2014: SJDM>

(54) **you're not buying** dodgy tablets off ebay or anything? <BNC2014: SPZR>

(1) there you are cos er <anon type="name" nameType="m"/> **wasn't using** it then? <BNC2014: S263>

These notable differences in the use of both negated and question-form progressives in textbook conversation as compared to naturally occurring interactions between native speakers may be associated with Carter's (1998: 47) claim that many coursebook dialogues portray:

> a 'can do' society, in which interaction is generally smooth and problem-free, the speakers co-operate with each other politely, […] and the questions and answers are sequenced rather in the manner of a quiz show or court-room interrogation.

Similarly, in his comparison of textbook vs. authentic service encounter dialogues, Gilmore (2004; see 2.2.4) also observed that, unlike textbook conversations, naturally occurring dialogues contain a lot of repetition, backtracking, misunderstandings and clarifications. Indeed, excerpts such as (51) seem to suggest that, at least in the use of questions in textbook dialogues, little has changed over the past decades.

Progressives in interrogative contexts are less frequent in both the Textbook Fiction subcorpus and the Youth Fiction reference corpus (7.25% and 8.44% of progressive concordance lines, respectively). No significant difference between the two corpora was found concerning this variable ($\chi^2(1) = 1.3055$, $p = 0.25$).

4.3.2.5 Voice

Römer (2005: 3) did not include the active-passive paradigm in her study on the progressive in contemporary ENL and Textbook English; therefore, to the present author's best knowledge, this is the first investigation of progressive passives in EFL textbook language.

Overall, Table 21 shows that passive progressives are very rare across all four sub(corpora) under study. Whilst the raw figures may imply that progressive passives are slightly underrepresented in the TEC subcorpora, statistical tests suggest that there is no meaningful association between the corpora and the proportions of active/passive voice progressives.

Table 21: Voice of progressive constructions

	Active	**Passive**	**Unclear**
Textbook Conversation	99.42% (2409)	0.58% (14)	0.00% (0)
Spoken BNC2014 sample	99.01% (2399)	0.95% (23)	0.04% (1)
Textbook Fiction	99.41% (1508)	0.59% (9)	0.00% (0)
Youth Fiction sample	98.22% (1490)	1.78% (27)	0.00% (0)

Based on his comparison of LOB and FLOB, Smith (2005: Chapter 6) reports that progressive passives in the present tense have seen a dramatic increase in contemporary British English. He points out that this increase is all the more striking because, over the same period, a significant decrease in non-progressive passives has also been observed (Smith 2005: 138). However, Smith's analysis refers to published writing genres only. Due to the very low raw figures in the LOB and FLOB, he also compares the frequencies of progressive passives across the different genres of the BNC1994 and the ICE-GB and his comparison clearly shows that the lowest relative frequencies are found in conversation, with 29 and 113 progressive passives per million words (pmw) in the Spoken BNC1994 and the ICE-GB conversation subcorpora respectively, compared to over 400 pmw in newspaper writing, broadcast news and institutional documents.

Overall, Table 21 confirms these low frequencies of passive progressives for the conversation register. However, of these few occurrences, nearly a third ($n = 11$) were not in the present progressive, the tense for which the substantial increase in passive use has been observed, but rather in the past form (55)–(56).

(55) The man **was being chased** by a police officer because he'd stolen a gold necklace from a jewellery shop. <TEC: New Green Line 5>

(56) we just started talking about you and then we realised that we **were being recorded** <BNC2014: SP2X>

A qualitative analysis of the few passive progressives observed in either conversation corpora reveals a notable difference in the degree of colloquialism. Indeed, the Spoken BNC2014 sample features a total of six GET-passives (57) and one HAVE-passive in the progressive (58), whereas these are entirely absent from the Textbook Conversation subcorpus.

(57) how much **am I getting paid**? <BNC2014: SA7J>

(58) cos I know some women they're find it a bit difficult mm to cope with the fact that **they're having their lady bits taken away** <BNC2014: S28F>

These alternative passive forms are generally considered more colloquial than progressive passives formed with BE (Smith 2005: 125). Indeed, the BE-passives featured in Textbook Conversation seem more typical of written genres than spontaneous conversation (59)–(62).

(59) Countries that still use leaded petrol **are being urged** to stop using it […] <TEC: Solutions Intermediate Plus>

(60) Well, Amy, if you must know - It isn't only to watch sport. He's going to run In a triathlon. And **he's being sponsored**, you know, for charity. <TEC: English in Mind 4>

(61) I push myself to the limit, and I'm usually covered in bruises! But I do try to avoid injury, though it's difficult if you **are being taught** a new move. <TEC: Solutions Intermediate>

(62) today in South Africa about one third of our water **is being wasted** because of old pipes and insufficient infrastructure. <TEC: New Mission 2de>

4.3.3 Functions of the progressive

4.3.3.1 Time reference

Fig. 9 shows considerable distributional differences in the time period progressives in Textbook Conversation and the Spoken BNC2014 sample refer to. The largest numerical differences are found in the categories 'past' and 'present' time reference but considering the distribution of tense forms reported in 4.3.2.1, this is to be expected. The high number of 'unclear' cases in the Spoken BNC2014 is likely due to the high degree of implicitness typical of spontaneous conversation between speakers who share a common environment and previous knowledge.

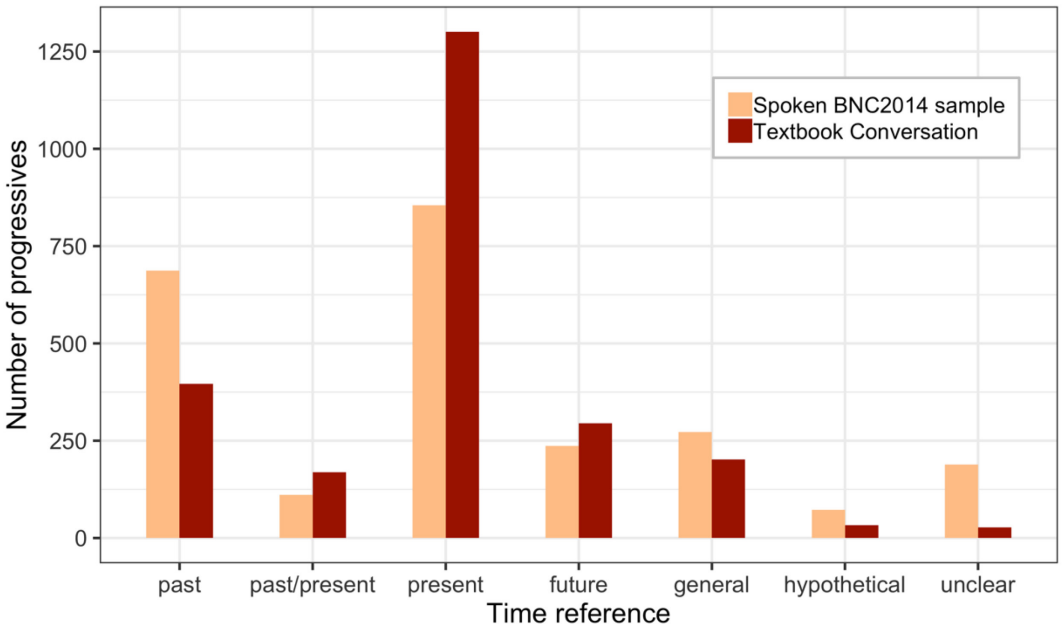


Fig. 9: Time period reference for the progressives in Textbook Conversation and the Spoken BNC2014 sample

The category 'past/present' time reference refers to events and situations which started in the past and are ongoing (63). Fig. 9 suggests that this time reference is over-represented in Textbook Conversation. Like the differences observed in the time reference categories 'past' and 'present', this is also due to the over-representation of present perfect forms in Textbook Conversation noted in 4.3.2.1.

(63) Come on! **We've been waiting** for fifteen minutes already!
<TEC: Green Line 4>

The differences observed in the time reference categories 'future', 'general validity', and 'hypothetical' correspond to differences previously reported by Römer (2005: 257) in her comparison of "school" and "real" English, though she groups the latter two time references in a single category: 'present/future ("indeterminate") time reference'. For this category, Römer reports a considerable difference in proportional usage (15.99% and 14.38% in spoken English versus 3.58% and 2.78% in school textbooks). At the same time, she also finds that future time reference is over-represented in EFL textbook dialogues (31.60% and 31.65% compared to 18.55% and 15.68% in her reference spoken English corpora).

Whilst interesting in its own right, Römer's presentation of the results makes it difficult to tell which categories make a relevant contribution to the overall distributional difference. To remedy this in the context of the present study, an association plot (Meyer et al. 2003) of the data presented in Fig. 9 was produced (see Fig. 10). In Fig. 10, the colour of the shading shows that, whilst the overall time reference distributions are statistically different at $p < 0.001$ (Cramér's $V = 0.26$), the category 'future' time reference does not make a statistically significant contribution to this distributional difference. By contrast, the proportions of

progressives found in the categories 'general validity' (64)–(65) and 'hypothetical' reference (66)–(67), are significantly associated with the corpora/variety so that Fig. 10 suggests a genuine under-representation of these functions in Textbook Conversation.

(64) her mum's not skinny either like her mum's a middle aged woman she**'s carrying** a bit of weight around her stomach but she doesn't she's not fat but she's not skinny <BNC2014: SAMQ>

(65) and when she went for the entry exam the guy said right at this point you know those of you have passed need to do this and this to get ready for your fitness then he said you should be aiming to run a mile and a half in sixteen minutes well she**'s running** it in fifteen and a half <BNC2014: SBYQ>

(66) no I think she **would** still **be doing** it
I think she probably would because I mean it would the same if I came with everybody it would be even worse if you are all up there and I **was doing** doing nothing it would be even worse if you are all up there and I **was doing** nothing <BNC2014: S3M4>

(67) but when you**'re earning** what I**'m earning** there's no way I**'m paying** off five hundred pound is there <BNC2014: SJLT>

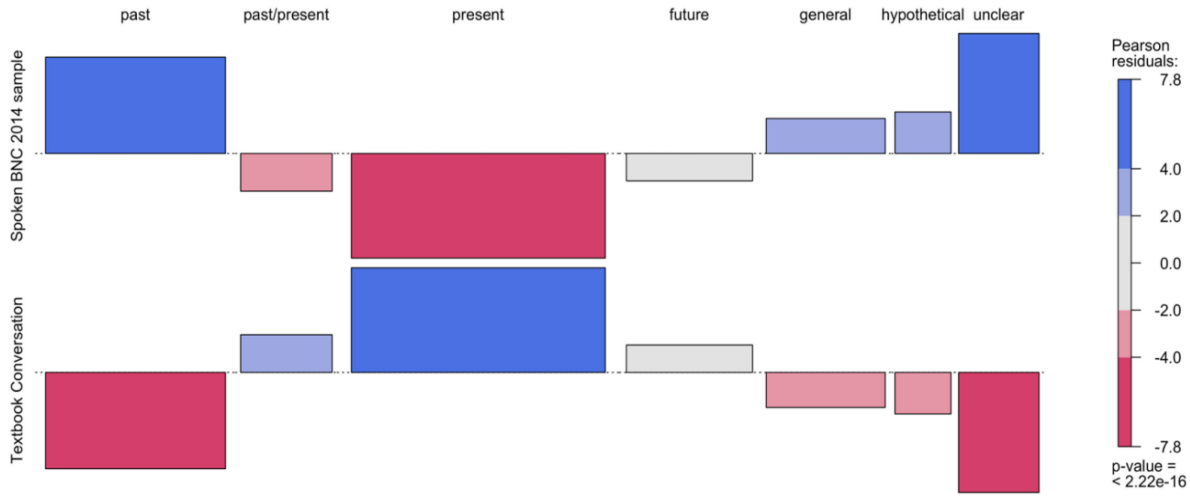


Fig. 10: Association plot of the distribution of time references of progressives in Textbook Conversation and the Spoken BNC2014 sample

To understand the pedagogical decisions made by the textbook authors, it is important to take a differentiated look at the distribution of time references of progressives at each Textbook level (see Fig. 11). The proficiency levels of the textbooks of the TEC are labelled from A to E – A corresponding to first year secondary school textbooks (see Table 3 for details).

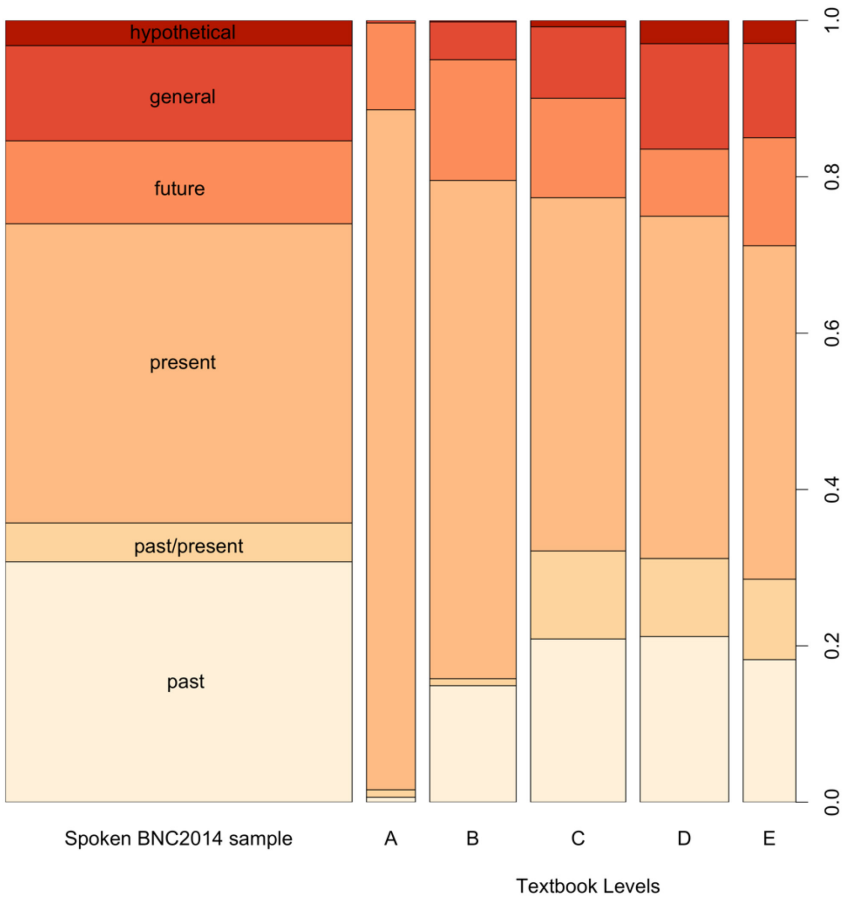


Fig. 11: Distribution of time references of progressives in the Spoken BNC2014 sample and in Textbook Conversation for each textbook level[25]

What can be clearly seen in Fig. 11 is the strong dominance of the 'present' and 'future' time references in the beginner textbooks, which is to be expected since, in Fig. 4, these textbooks were shown to feature almost exclusively present tense form progressives. Moreover, it is evident that the present perfect form and its prototypical past/present time reference are not introduced until the third year of EFL teaching (here: level C).

More importantly, Fig. 11 reveals that the time reference distributions of the intermediate and advanced textbooks (levels C to E) are, in fact, relatively close to that observed in the Spoken BNC2014 sample. Whilst an independence χ^2 test of this subset of the data nevertheless showed a significant association between time reference distribution and naturally occurring/textbook conversation ($\chi^2(5) = 88.371$, $p < 0.001$), the very modest strength of association (Cramér's $V = 0.15$) confirms the fact that the distribution of time references of the more advanced textbooks does indeed match relatively well that of naturally occurring conversation. What is more, the association plot between advanced Textbook Conversation – levels C to E only – and the Spoken BNC2014 sample (see Fig. 12) shows that the most important differences are found in the categories 'unclear', which is of little pedagogical

[25] Note that for ease of comparability, 'unclear' concordance lines were removed from this mosaic plot.

relevance, and 'past/present'. This latter divergence will be further explored in the following section.

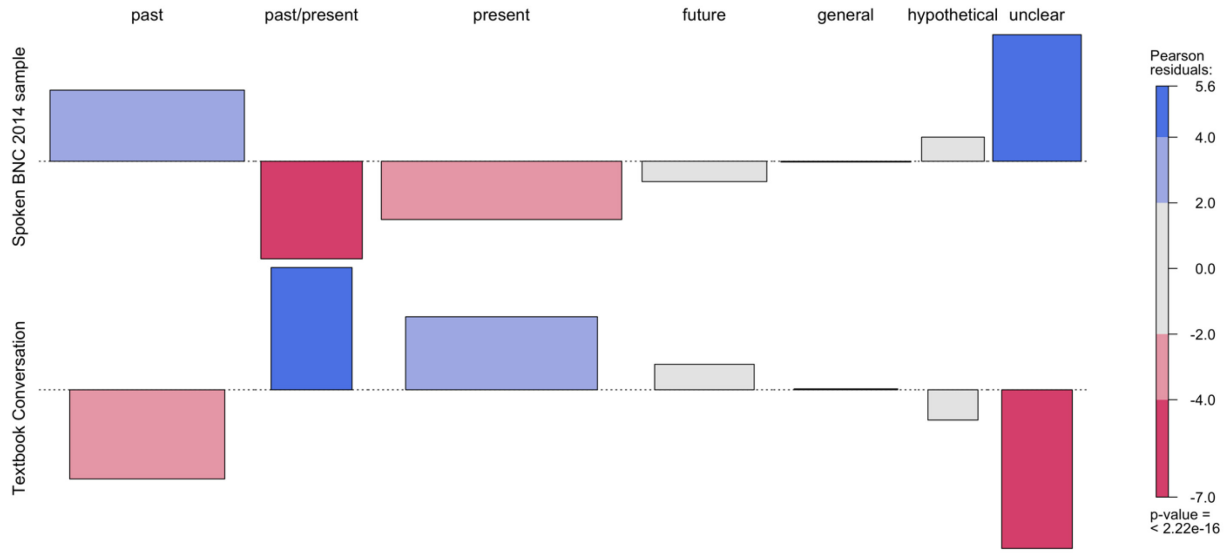


Fig. 12: Association plot of progressive time reference between Level C-E Textbook Conversation and the Spoken BNC2014 sample

The Textbook Fiction vs. Youth Fiction comparison revealed no notable differences in the distributions of time reference of progressives. Whilst significant overall ($\chi^2(6) = 41.701$, $p < 0.001$, Cramér's $V = 0.12$), the only category that makes a significant contribution to this difference is the 'present' time reference which, as described above for Textbook Conversation, is due to the "weight" of the beginner textbooks in the subcorpus. Indeed, an association plot comparing the progressives in the narrative texts of the three more advanced textbooks in each textbook series with those of the Youth Fiction corpus yielded no significant difference between the time reference distributions ($\chi^2(5) = 11.662$, $p = 0.04$).

4.3.3.2 Time reference, tense and modality

Fig. 13 displays a cross tabulation of language variety (top and bottom plots), progressive tense form (vertical subdivisions) and time reference (horizontal subdivisions). It enables a more fine-grained comparison of the functional usage of the different progressive tense forms across Textbook and naturally occurring ENL conversation.

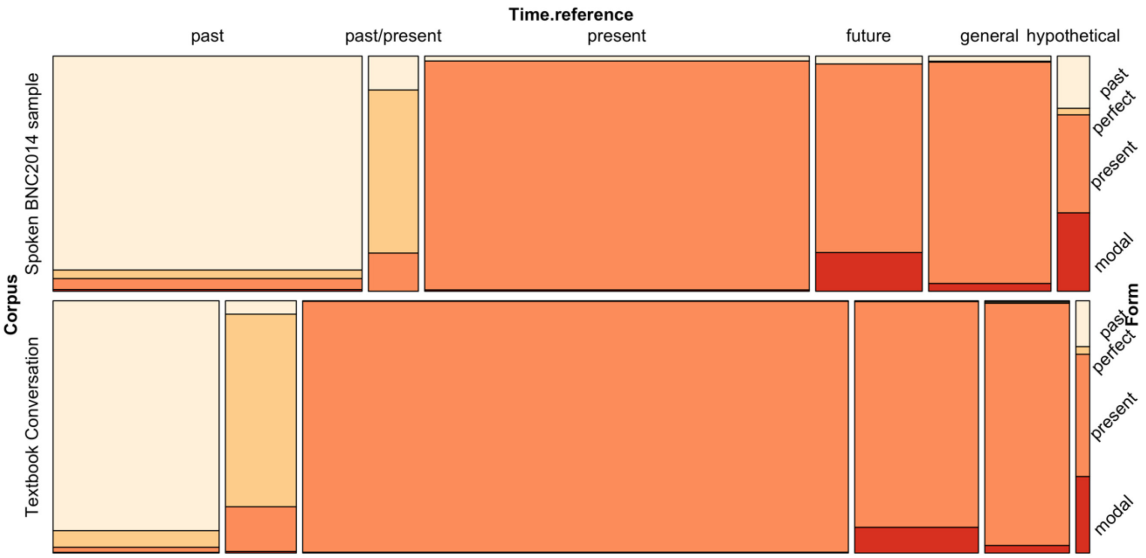


Fig. 13: Mosaic plot displaying the cross-tabulation of forms and time references of progressives in Textbook Conversation and the Spoken BNC2014 sample (progressives with 'unclear' time references were removed for this analysis).

It shows that, in Textbook Conversation, perfect forms are used to refer to both events that started in the past and are still ongoing and past events which preceded another past event. This is due to the fact that a number of textbook series introduce the past perfect progressive form in the more advanced EFL textbooks (68)–(70) even though this form is very rare in naturally occurring conversation (no single occurrence among the 2,423 random progressive concordance lines from the Spoken BNC2014).

(68) I wonder if anyone **had been living** here before Verrazano came. <TEC: Green Line 4>

(69) His name was Mike. He was one of the owners of the hostel and **had been taking care** of the terrace plants. <TEC: New Green Line 5>

(70) I got home I and did an Internet search. It didn't take long. I **had been searching** for less than a minute when I found the answer: the artist I was looking for was JR. <TEC: Access 5>

It may be noted that the three past perfect progressive examples from the Textbook Conversation sub-corpus (68)–(70) are all taken from German textbook series. Strikingly, a CQL search for past progressives across all textbook registers confirmed that this form is particularly frequent in German textbooks (accounting for 59.67% of all occurrences in the full Textbook English Corpus).

Fig. 13 also reveals that British English native speakers are more likely to resort to *modal + progressive* constructions (foremost WILL + BE + *-ing*) to refer to (possible) future events (71)–(73) than the interlocutors of the Textbook dialogues.

(71) it was a special little treat
I don't think I'**ll be going** back again any time soon though unfortunately

might do that for like graduation or something that'd be nice
<BNC2014: SEM7>

(72) even if it turns out to be shit at least it's a plan
yeah
and at least it means that I**'ll be working** towards something so
<BNC2014: SYTD>

(73) Nunnington Hall **will be reopening** soon cos it
will it?
closes over winter <BNC2014: S4QF>

The few occurrences of past tense forms with future time reference from the Spoken BNC2014 are due to reported speech and thought (74).

(74) I've got to go to Waitrose to pick up the turkey when **are you working**
and stuff like that?
I thought **you were getting** turkey Thursday
no **getting** the turkey Wednesday after work
where you gonna put it? <BNC2014: SNK6>

In sum, though Fig. 13 shows fairly similar distributions of tense forms cross-tabulated by time reference in Textbook Conversation and the Spoken BNC2014, a few notable differences have been highlighted above. By contrast, the equivalent analysis for Textbook Fiction and Youth Fiction did not point to any further potentially relevant differences worth investigating in more depth. The only substantially different tense form distribution is found in the time reference category 'past/present' which, as above for Textbook Conversation, is made up exclusively of present perfect progressive forms in the TEC subcorpus, whereas Youth Fiction also includes past progressives in this time reference category.

4.3.3.3 (Non-)continuousness

Many EFL grammars, and indeed the three textbook series representing the Spanish Textbook subcorpus (*Achievers, English in Mind* and *Solutions*), refer to the progressive construction as "the continuous". It may therefore come as a surprise that a considerable proportion of progressives in the two reference corpora (21.38% in the Spoken BNC2014 sample and 15.53% in the Youth Fiction sample) actually refers to 'non-continuous' actions (75).

(75) so that's why I kind of like keep my answers short when she's when **she's asking** me how I am mm <BNC2014: S954>

By contrast, in the two Textbook English registers under study, progressives are less frequently used for 'non-continuous' actions (see Table 22). Table 22 shows that the proportion of 'non-continuous' progressives in the Spoken BNC2014 is almost twice as high as in Textbook Conversation. The null hypothesis that there is no association between naturally occurring/textbook conversation and the proportion of 'non-continuous' progressives can be rejected for both the conversation register

($\chi^2(1) = 131.04$, $p < 0.001$, $\varphi = 0.14$) and the Textbook Fiction/Youth Fiction pair ($\chi^2(1) = 26.76$, $p < 0.001$, $\varphi = 0.09$).[26]

Table 22: The continuousness function of progressives in the two textbook registers and reference corpora

	Continuous	*Non-continuous*	*Unclear*
Textbook Conversation	88.40% (2412)	11.35% (275)	0.25% (6)
Spoken BNC2014 sample	75.57% (1831)	21.38% (518)	3.05% (74)
Textbook Fiction	89.78% (1362)	9.16% (139)	1.05% (16)
Youth Fiction sample	83.78% (1271)	15.43% (234)	0.79% (12)

A lexico-grammatical approach to the continuous/non-continuous paradigm offers further insights into the differences between progressive usage in naturally occurring ENL conversation and Textbook Conversation. Table 23 lists the lemmas that occur most frequently with a 'non-continuous' function in either corpora. Some frequency ranks in the Textbook Conversation column contain more than one lemma whenever these have the same number of occurrences in a 'non-continuous' context. These raw numbers of occurrences (out of the total of 2,423 progressives investigated in each of the two datasets) is indicated in brackets. Additionally, the percentages in brackets refer to the proportions of 'non-continuous' usage for each lemma in the corresponding corpus.[27]

[26] Concordance lines for which it was not possible to reasonably guess whether a progressive was referring to a continuous or non-continuous action (i.e., 'unclear') were not taken into account for both these χ^2 tests.

[27] Note that a number of lemmas absent from the Spoken BNC2014 rank list, such as ASK, BUY, LEAVE AND pick up, display similarly high proportional usage of the 'non-continuous' function in naturally occurring conversation, but are overall comparatively far less frequent in progressive constructions than in Textbook Conversation, hence why they do not appear in this rank table.

Table 23: Most frequent lemmas occurring in the progressive with a 'non-continuous' function in Textbook Conversation and the Spoken BNC2014 (total occurrences in 'non-continuous' function) (as a percentage of all progressive occurrences)

Rank	Spoken BNC2014 sample	Textbook Conversation
1	SAY (77) (51%)	GO (52) (32%)
2	GO (76) (48%)	COME (16) (30%)
3	GET (36) (49%)	ASK (9) (60%) / LEAVE (60%) / SAY (31%) / START (64%)
4	COME (31) (60%)	BUY (8) (67%)
5	HAVE (21) (34%)	TAKE (7) (21%)
6	TAKE (17) (65%)	GET (6) (14%) / JOKE (55%) / MEET (46%) / TELL (32%)
7	TELL (14) (52%)	CALL (4) (33%) / DO (2%) / HAVE (7%) / KID (31%) / PICK UP (100%) / SEND (100%)

Table 23 indicates that, in the Spoken BNC2014 sample, SAY is the verb lemma most frequently associated with the non-continuous function in the progressive (see Table 23). In this sample, every other progressive occurrence of SAY refers to a 'non-continuous action' (76)–(77), whereas two thirds of progressive SAY occurrences in Textbook Conversation actually refer to 'continuous' actions, ongoing at the time of speaking, e.g., (79).

(76) but I read the text and the text **is** basically **saying** that there's three parts to people there's the emotional which is what you are when you are in your mother's womb <BNC2014: SEKZ>

(77) paid off the builders yeah well they**'re saying** the the if they do expand the A14 it will be paid for the tolls <BNC2014: S8LS>

(78) deal with the estate agents deal with the plans like he **was saying** you've got to have a look at the whole planning act <BNC2014: SG56>

(79) A: Did they enjoy your talk?
B: I don't think so. I don't think they understand a word of what I said.
A: What**'s he saying** now?
B: I don't know. It's too noisy in here. <TEC: English in Mind 4>

A number of other verb lemma, such as ASK, GET, HAVE, TAKE and TELL, follow a similar trend: in the naturally occurring conversation data, progressive forms frequently express non-continuousness (80)–(81), whilst in the textbook dialogues, the focus is still refer to something procedural (82) or to ongoing actions (83).

(80) the last one I saw was erm having to pay tax on tampons oh I haven't seen that one it's it**'s asking** the government to get rid of tax on tampons cos it or it's already tax free on crocodile meat <BNC2014: S6MQ>

(81) if you don't do this it's me I'**ll be having** a go at you because it's my job to stop you getting diabetes <BNC2014: SGMT>

(82) So, let's get started! I'm sure you**'ve been asking** yourselves: What makes Seattle so special? <TEC: New Green Line 4>

(83) It's twelve o'clock at night in Edinburgh and we**'re having** a great time, it's incredibly noisy here! <TEC: Achievers A>

4.3.3.4 Repeatedness

Table 24 summarises the results of the repeatedness function analysis. It shows that progressives which express repeated actions or states are under-represented in Textbook Conversation ($\chi^2(1) = 30.52$, $p < 0.001$, $\varphi = 0.082$)[28]. In contrast, the proportion of progressives which refer to repeated and non-repeated actions in Textbook Fiction is not significantly different to that of the Youth Fiction sample ($\chi^2(1) = 12.61$, $p = 0.21$).

Table 24: The repeatedness function of progressives in the two textbook registers and reference corpora as a percentage of all progressives (and in absolute figures)

	Non-repeated	*Repeated*	*Unclear*
Textbook Conversation	82.58% (2,001)	16.26% (394)	1.16% (28)
Spoken BNC2014 sample	69.29% (1,679)	20.68% (501)	10.03% (243)
Textbook Fiction	89.58% (1,359)	8.77% (133)	1.65% (25)
Youth Fiction sample	86.49% (1,312)	9.95% (151)	3.56% (54)

In the Spoken BNC2014 sample, a number of verb lemmas in the progressive very frequently express repeatedness. Among the top 30 verb lemmas most frequently observed in the progressive, these include PAY (84%), USE (60%), WORK (51%), WATCH (40%) and PUT (40%). The highest proportional use of the repeated function in Textbook Conversation is found with the lemma USE but, with 38%, at an appreciably lower rate than in the Spoken BNC. The second highest is WORK, with 36%.

Although the figures may not sound particularly alarming, as previously warned by Römer (2005: 284), little exposure to this core function could lead to genuine misunderstandings in the foreign language. For instance, EFL learners may not recognise the repeated nature of these events in the following extracts from the Spoken BNC2014, e.g., (84)–(85).

(84) oh they've taken the one that you put er compostables in and garden rubbish so that means people **are putting** them in the black bins <BNC2014: SRWD>

(85) **she was trying** to multitask everything that was going on in her life mm she's she was like getting chest pains and stuff from it <BNC2014: ST47>

[28] This test of independence and the next were performed without the concordance lines labelled 'unclear' for these features, though, for repeatedness, it is worth noting that these make up a non-negligible 10% of the Spoken BNC2014 data.

4.3.3.5 Additional functions of the progressives

Table 25 summarises the number of progressives coded with additional functions following an adaptation of Römer's (2005) functional categories (described in 4.2.2.5). Of those, the null hypothesis that there is no association between the prevalence of any one of these additional functions and the source of the data (i.e., Textbook vs. ENL reference data) can be rejected for both registers in the case of 'framing' (Conversation: χ^2 (1) = 19.56, $p < 0.001$, $\varphi = 0.24$; Fiction: χ^2 (1) = 26.13, $p < 0.001$, $\varphi = 0.25$) and 'gradual change/development' (Conversation: $\chi^2(1) = 9.55$, $p < 0.005$, $\varphi = 0.17$; Fiction: $\chi^2(1) = 13.84$, $p < 0.001$, $\varphi = 0.18$). No significant differences are observed for the other two categories.

Table 25: Percentages of progressives within each (sub)corpus coded for additional functions (raw figures are given in brackets)

	Emphasis /shock	*Framing*	*Gradual change/ development*	*Politeness/ softening*
Textbook Conversation	1.53% (37)	4.04% (98)	1.94% (47)	0.21% (5)
Spoken BNC2014	1.65% (40)	1.86% (45)	2.72% (66)	0.33% (8)
Textbook Fiction	0.79% (12)	9.49% (144)	3.36% (51)	0.13% (2)
Youth Fiction	1.32% (20)	5.21% (79)	3.36% (82)	0.13% (2)

The considerable over-representation of progressives in a 'framing' context in textbook registers undoubtedly stems from the fact that the 'framing' function is central to description of the progressive in the textbooks' grammar sections (see Römer 2005: 241). It is given particular emphasis when the past progressive is introduced, as a way to contrast the past progressive and the past simple. The analysis of the progressive concordance lines shows that this 'framing' use of the past progressive is modelled in many of the textbook dialogues and narrative texts (86)–(89). In contrast, the 'framing' function is less frequently featured in ENL fiction and is very rare in naturally occurring conversation (see Table 25).

(86) While he **was walking** through customs his clothes fell out of his bag. <TEC: Green Line 3>

(87) While they **were listening**, he thought about his parents. <TEC: New Green Line 4>

(88) Sorry I missed your call yesterday. I **was having** my dance class when you rang. <TEC: Hi There 5^{e}>

(89) While the Prince and his men **were riding** through the forest, they heard loud barking […] <TEC: English in Mind 3>

4.3.4 Lexical and semantic aspects

4.3.4.1 Co-varying Collexeme Analyses of 'Verb lemma' + 'Variety'

In order not to conflate the effects of language variety (i.e., Textbook English vs. naturally occurring ENL) with register, two separate CovCAs comparing how each verb lemma is attracted to or repelled by the progressive construction in each variety, one for each register under study, were conducted. Accordingly, Table 26 presents a ranking of the top twenty verb lemmas whose use in the progressive construction is most strongly associated with either Textbook Conversation or the Spoken BNC2014, whilst Table 27 lists the equivalent results for Textbook Fiction and the Youth Fiction corpus. Both tables indicate the raw and expected frequencies of the top-ranking verb lemmas in each corpus and the verb's collostructional strength (CL), which, here, represent the degree of association between a verb lemma and a variety.

The top-ranking collexemes associated with the progressive constructions in the two conversation samples are displayed in Table 26. The table features some remarkably high CL values, thus pointing to major differences in the lemmas featured in the progressive in Textbook Conversation as compared to the Spoken BNC2014.

Table 26: Top 20 Co-varying Collexemes for 'Verb' + 'Variety' (Spoken BNC2014 vs. Textbook Conversation)

Attraction to Textbook Conversation				Attraction to Spoken BNC2014 sample			
Verb	**Observed Freq.**	**Expected Freq.**	**CL**	**Verb**	**Observed Freq.**	**Expected Freq.**	**CL**
WEAR	71	42.5	10.06	SAY	153	91	21.84
LOOK	92	61	8.26	THINK	101	64	11.08
WAIT	50	33.5	4.52	RECORD	15	8	3.59
PLAY	56	40	3.69	TRY	98	76.5	3.58
STUDY	18	10.5	3.14	GET	75	59.5	2.59
WRITE	18	10.5	3.14	PAY	19	12	2.49
VISIT	10	5	3.01	BE	34	25	2.13
SLEEP	14	8	2.69	GO OFF	6	3	1.81
LISTEN	28	19	2.61	SWEAT	5	2.5	1.51
FEEL	34	24	2.58	PUT	15	10.5	1.41
SMILE	11	6	2.50	LOOK FOR	22	16.5	1.40
MEET	13	7.5	2.44	STRUGGLE	6	3.5	1.21
DANCE	8	4	2.41	CONSIDER	4	2	1.20
PREPARE	8	4	2.41	CUT	4	2	1.20
ENJOY	17	11	2.08	DEAL	4	2	1.20
SWIM	9	5	1.97	EARN	4	2	1.20
WALK	27	19.5	1.94	GO DOWN	4	2	1.20
STAND	19	13	1.85	LOOK UP	4	2	1.20
STAY	25	18	1.85	PICK	4	2	1.20
HURT	5	2.5	1.51	MAKE	42	35.5	1.12
PASS	5	2.5	1.51	GIVE	10	7	1.05

The majority of the top collexemes on the Textbook Conversation side belong to either the semantic domain of activity verbs (PLAY, VISIT, STUDY, WRITE, LISTEN, DANCE, SWIM, WALK etc.) or can be qualified as descriptive verbs (WEAR, SMILE, FEEL etc.). The first domain corresponds to a highly prototypical use of the progressive, often the very first one to be introduced in EFL textbooks – as illustrated with extracts from the grammar sections of two beginner textbooks in (90) and (91).

(90) I **am listening** to the radio. You **are doing** your homework. He **is looking** at a magazine. She **is watching** a game on TV. It **is starting** right now. We **are making** a poster. You **are having** a nice time. They **are winning** now.
Du bildest die Verlaufsform so: Form von to be (am, is, are) Vollverb (Infinitiv ohne to) ing-Endung: She **is watching** TV.
Mit der Verlaufsform der Gegenwart kannst du ausdrücken, dass jemand gerade dabei ist, etwas zu tun. Du kannst das present progressive auch verwenden, wenn du sagen möchtest, dass ein Vorgang gerade abläuft und noch nicht abgeschlossen ist. <TEC: Green Line 1>

(91) To talk about things that are happening now. E.g. **I'm reading** on the roof. <TEC: English in Mind Starter>

The second domain is less commonly the subject of explicit instruction, though one French textbook series does introduce the progressive with the following function and example sentence:

(92) Description of an activity: *The ghosts* ***are playing*** *football.* <TEC: Piece of Cake 6e>

Moreover, picture description tasks are heavily relied on in EFL textbooks as a means of enabling learners to practise using the progressive. Indeed, over the course of the manual annotation process, it was also noted that some 80 occurrences of the present progressive in the Textbook Conversation subcorpus were part of picture or film descriptions, e.g., (93)–(94). In many instances, these sentences are acting as a model for picture/film description tasks set at a later stage in the textbooks (94). This task-effect likely contributes to the high CL scores of verbs such as WEAR, SMILE, FEEL, ENJOY and LOOK in Textbook Conversation.

(93) On the cover, there is a big woman wearing goggles, a type of protection glasses, and a shield. She **is wearing** pins for victory and blood donations on her jacket. She **is** also **wearing** overalls and there is a massive riveting gun on her knees. […] But she's also feminine because she**'s wearing** lipstick. […] She**'s wearing** nail polish and has curly red hair. <TEC: Piece of Cake 3e>

(94) The photo shows five people in a dinghy. They're white-water rafting on a river. In the top left corner of the photo, there is a man leaning out of the boat. In the foreground, there's a lot of water. The man at the back of the dinghy is trying to guide the dinghy with his paddle. The other four people **aren't helping** very much. It looks as if they're new to it. They'**re** all **wearing** the same life jackets and helmets. I imagine they**'re doing** this as a holiday activity. <TEC: Pre-intermediate Solutions>

EFL learners are known to frequently overuse BE in the progressive and, indeed, this is considered to be a typical error in many EFL learning contexts. Thus, textbooks frequently include BE in their lists of stative verbs that are normally not to be used in the progressive. It therefore comes as no surprise that, in the progressive, BE is more strongly associated with the Spoken BNC2014 (CL = 2.13), whilst it is strongly repelled by Textbook Conversation. Across the entire Spoken BNC2014 some of the top adjective collocates immediately to the right of the verb form *being* include *able, funny, sick, serious, silly, rude* and *gay*. Of these, only *being silly* is featured in Textbook Conversation ($n = 3$), thus it may be postulated that, in spite of being featured in many bestselling EFL grammars (e.g., Murphy et al. 2009: 8), this specific use of the progressive BE + *adjective* (usually describing human behaviour) is in fact underrepresented in Textbook Conversation.

Interestingly, the second highest collexeme in the Textbook Conversation ranking in Table 26 is LOOK, a verb on which Römer (2006b) dedicated an entire article, hereby pointing to some striking differences between the functions of LOOK in the progressive in textbook dialogues and in natural spoken English. Römer (2006b) identified LOOK AT as an overused verb-preposition collocation in the dialogues of an older version of the German textbook series *Green Line*. By teasing out the different prepositional and phrasal verbs associated with LOOK from this analysis (see 4.2.3.1), it can be concluded that this overuse of LOOK in the progressive in Textbook Conversation is most strongly linked to the senses SEE and APPEAR, as opposed to LOOK FOR (= SEARCH), whose progressive form is more attracted to the Spoken BNC2014, and LOOK AFTER, LOOK FORWARD, etc., that do not feature in Table 26 at all.

The top-ranking collexemes from the Spoken BNC2014 also feature a number of activity verbs, but these are subject to corpus topic effects and thus quite different in terms of semantic domains. Unsurprisingly, whilst many of the activities from the Textbook Conversation side of Table 26 are typical school or teenager activities (e.g., PLAY, SLEEP, STUDY, SWIM, WRITE, VISIT, DANCE, LISTEN), the Spoken BNC2014 side features more activities from the world of work and, more generally, adulthood (e.g., PAY, DEAL, EARN). Broadly speaking, it would appear that the world portrayed in school EFL textbooks is seemingly a largely positive, upbeat affair whilst the naturally occurring conversations from the Spoken BNC2014 also feature more nuanced progressive collexemes such as TRY, SWEAT and STRUGGLE.

The fact that the verb lemma RECORD is the third highest ranking collexeme for the Spoken BNC2014 is clearly an artefact of the data collection process. Indeed, the new Spoken BNC was created by asking members of the public to record their conversations on their own smartphones (Love et al. 2017: 329). This unusual set-up prompted regular remarks by the interlocutors of the conversations featuring RECORD in the progressive such as (95) and (96). At the end of the day, the high CL value of RECORD serves as a reminder of the inherent difficulty of obtaining genuinely "authentic" or "naturally occurring" speech data.

(95) oh you're still oh er you**'re recording**? oh you'**re not recording** this? yeah yeah it's perfect perfect talk <BNC2014: SAUR>

(96) oh no no don't don't that **was** it **recording** when I told you to F off? yeah but it's fine it doesn't matter <BNC2014: SAZX>

Looking at the highest ranking collexemes from the Spoken BNC2014 half of Table 26, another semantic domain emerges: SAY, THINK, CONSIDER, and in many cases, TRY all function as communication or discourse verbs. As illustrated in (97)–(99), these verbs frequently occur in relatively fixed phrasemes which help to structure discourse. We may speculate that if learners are not exposed to these phrasemes via their

textbooks, they may have difficulty interacting in genuine conversation settings where they are so frequent.

(97) cos **she was saying** um oh I just want to be on the boat <BNC2014: SBYQ>

(98) I suppose we might yeah that's **I was thinking** just go for coffee or something somewhere <BNC2014: SCWC>

(99) yeah like **I'm trying to say** really is that if had plenty of time <BNC2014: SBG4>

Table 27 lists the verb lemmas most strongly associated with the progressive in Textbook Fiction texts as compared to the Youth Fiction corpus. When interpreting these results, it is important to remember that the CL values from Table 26 cannot be compared to those in Table 27 because the total number of progressives investigated in each table differs and, since CL is calculated on the basis of *p*-values, it correlates with sample size. This notwithstanding, the results presented in Table 27 still suggest that there are considerably fewer major discrepancies in the verbs featured in the progressive in Textbook Fiction and the reference Youth Fiction corpus.

Table 27: Top 10 Co-varying Collexemes for *'Verb' + 'Variety'* (Textbook Fiction vs. Youth Fiction sampled)

Attraction to Textbook Fiction				Attraction to Youth Fiction sample			
Verb	**Obs. Freq.**	**Exp. Freq.**	**CL**	**Verb**	**Obs. Freq.**	**Exp. Freq.**	**CL**
WALK	39	21.5	7.9	BE	15	9	2.43
WAIT FOR	20	10	6.05	TAKE	12	7	2.2
SIT	61	42	4.8	HOLD	17	11	2.08
WEAR	37	25	3.37	LIE	18	12.5	1.67
VISIT	8	4	2.41	BEGIN	16	11	1.59
DO	82	69	1.84	HUNT	5	2.5	1.51
BURN	6	3	1.81	GIVE	9	5.5	1.49
EAT	11	7	1.55	GROW	6	3.5	1.21
LOOK FOR	20	14.5	1.52	FIGHT	4	2	1.20
START	14	9.5	1.50	WORRY	4	2	1.20
WRITE	9	5.5	1.49	STARE AT	7	4.5	1.05

It is tempting to identify similar topic effects to those described in the previous section on conversation; however, the raw frequencies of even the highest ranking collexemes make it clear that no significant discrepancies can be observed in the lexis of progressive verbs between Textbook Fiction and reference Youth Fiction. Nevertheless, this CovCA potentially points to two interesting phenomena. First, as in the conversation CovCA (see Table 26), BE is among the verbs in the progressive that are most strongly associated with the reference Youth Fiction corpus rather than with Textbook Fiction. Second, we turn to the highest ranking verb on the Textbook

Fiction side of the table: WALK. A qualitative analysis of its concordance lines reveals that it occurs in the progressive almost exclusively in framing contexts, e.g., (100)–(102). Framing is another prototypical function of the progressive, which has, however, been shown to be considerably less frequent in authentic usage than EFL grammars and textbooks would suggest (see 4.3.3.5 and Römer 2005: 267).

(100) Sam, Lucy, Leo and Maya **were walking** to the Broadway when someone suddenly shouted to them from a second-floor window. <TEC: Access 2>

(101) I **was walking** in the churchyard when suddenly a man jumped up from among the graves. <TEC: Join the Team 4[e]>

(102) While you **are walking** through the school building, you hear the principal's voice. <TEC: Green Line 4>

The following section presents two further CovCAs: zooming out from the lexical level to focus on the semantic domains distinctively associated with or repelled by Textbook Conversation and Textbook Fiction texts.

4.3.4.2 Co-varying Collexeme Analyses of 'Semantic domain' + 'Variety'

Before presenting the results of this second set of CovCAs on the semantic domains of verbs in the progressive, we briefly examine the overall distributions of semantic categories across the four corpora (see Fig. 14). Note that in both Fig. 14 and in the following CovCAs, the verbs categorised as 'other/unclear' have been excluded. This category accounts for the vast majority of polysemous verbs (e.g., GIVE which can be an 'activity' verb in *to give a present*, a 'mental' one in *to give proper consideration*, a 'causation' one in *to give cause,* etc.).

The distribution of semantic domains over the four corpora (see Fig. 14) broadly corresponds to what has been reported in prior research on the progressive: namely that the 'activity' domain accounts for around half of all progressive occurrences and that the 'aspectual' and 'causative' domains make up the smallest proportions of progressive occurrences (e.g., Deshors 2017: 273–274; Edwards 2014b: 173).

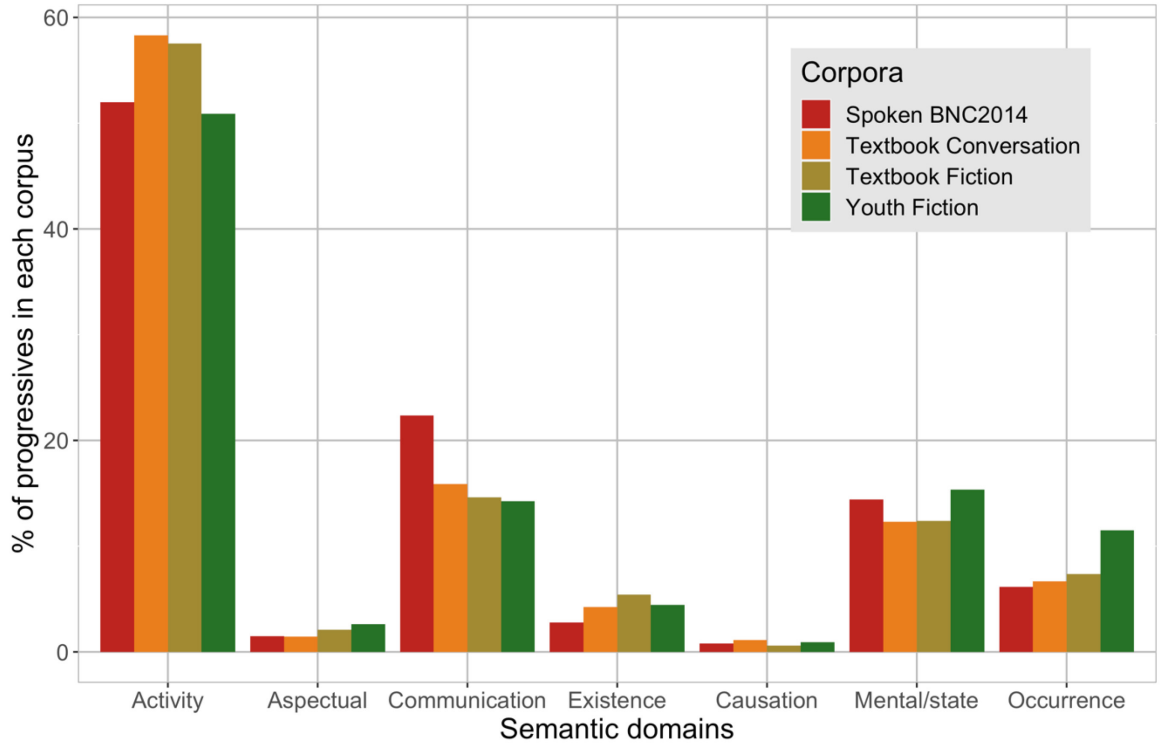


Fig. 14: Distribution of the semantic categories of verbs in the progressives across the four (sub)corpora (excluding polysemous verbs).

This second set of CovCAs makes use of the semantic domain annotation of the verbs of the progressive concordance lines under study in an attempt to identify domains that are specifically attracted to or repelled by the dialogues and narrative texts of EFL Textbooks (see Tables 28 and 30). As in 4.3.4.1, two separate analyses, one for each register, were conducted.

Table 28 presents the results of the semantic CovCA comparing the verb lemmas featured in the progressive in Textbook Conversation with those found in the Spoken BNC2014. The highest CL value in Table 28 is assigned to the 'communication' domain, which is strongly attracted to the Spoken BNC2014; in other words, strongly repelled by Textbook Conversation. This seems to confirm the phenomenon inferred from the previous lexis-based CovCAs where a number of 'communication' verbs such as SAY and CONSIDER ranked within the 20 progressive collexemes most repelled by Textbook Conversation compared to the Spoken BNC2014.

Table 28: Results of the CovCA 'Semantic domain' + 'Variety' (Textbook Conversation vs. reference Spoken BNC2014 sample)

Semantic domain	Attraction to	Obs. freq.	Exp. freq.	CL
Communication	Spoken BNC2014	370	315.41	5.99
Activity	Textbook Conversation	1003	949.67	3.90
Existence/relationship	Textbook Conversation	73	60.63	1.88
Mental/state	Spoken BNC2014	239	221.23	1.40
Causation	Textbook Conversation	19	16.30	0.66
Occurrence	Textbook Conversation	115	110.56	0.54
Aspectual	Spoken BNC2014	25	24.53	0.30

The difference with the previous lexis-based CovCA, however, is that this 'Semantic domain' + 'Variety' CovCA includes a total of 49 'communication' unique verb lemmas identified in the Spoken BNC2014 concordance samples. The most frequent ones are listed in Table 29. A total of 370 progressive concordance lines from the Spoken BNC2014 were annotated as having a verb within the semantic domain of 'communication'. Bearing this total in mind, we can see from Table 29 that two-thirds of these feature the verbs SAY and TALK. We can therefore speculate that this effect may actually be more lexical (i.e., largely driven by SAY, since TALK is not featured in Table 26) than semantic in nature.

Table 29: Eight most frequent 'communication' verbs in the Spoken BNC2014 progressive concordance sample (*n* = 370)

Communication verbs	Freq. in Spoken BNC2014 concordance sample
SAY	153
TALK	93
TELL	27
JOKE	8
SPEAK	7
TEACH	7
ASK	6
KID	6

In both 'Semantic domain' + 'Variety' CovCAs the highest ranking CL values for the textbook registers are those of the 'activity' domain. As hypothesised in 4.3.4.1 on the lexical CovCAs, this may be due to the prototypical functions of the progressive emphasised in the textbook grammar sections, as well as the description tasks intended to allow learners to practise using the progressive, both of which are modelled in a number of textbook dialogues.

Table 30: Results of the CovCA of *'Semantic domains' + 'Fiction Varieties'* (Textbook Fiction vs. Youth Fiction sample)

Corpus	Semantic domain	Obs. freq.	Exp. freq.	CL
Occurrence	Youth Fiction Sampled	135	110.42	3.47
Activity	Textbook Fiction	688	648.74	3.16
Mental	Youth Fiction Sampled	180	162.41	1.68
Existence/Relationship	Textbook Fiction	65	59.07	0.82
Aspectual	Youth Fiction Sampled	31	27.73	0.64
Causation	Youth Fiction Sampled	11	8.910	0.64
Communication	Textbook Fiction	175	172.66	0.38

Furthermore, the results summarised in Table 30 show that the semantic domain 'occurrence' is significantly more strongly attracted to the Youth Fiction reference corpus than to Textbook Fiction. The high lexical diversity of fictional writing means that this particular effect is undeniably semantic rather than lexical. Indeed, the most frequent 'occurrence' verb within the Youth Fiction concordance lines, go on, has a raw frequency of just 18 out of 135 concordance lines with 'occurrence' verbs. Further, qualitative examination of the Youth Fiction concordance lines featuring 'occurrence' verbs reveals that this attraction of the domain 'occurrence' to the reference corpus is the result of longer, more detailed descriptions in the novels of the Youth Fiction corpus. In contrast, the narrative texts in the textbooks focus more on the actions of characters and thus the activity domain is more strongly attracted to Textbook Fiction.

4.3.4.3 Correspondence Analysis (CA) of 'Semantic domain' + 'Variety' + 'Register'

This section presents the results of a simple binary CA with 'Semantic domain' and 'Corpus' as the contributing variables and 'Register' and 'Variety' as supplementary variables. The χ^2 test of independence between the two main variables returned 107.65 ($p < 0.001$), thereby indicating that there is indeed a significant relationship between the two variables. However, the total inertia of the contingency table is very low ($\Phi^2 = 0.019$). As a result, the data points are all relatively close to the point where the average 'Corpus' and 'Semantic domain' profiles coincide, which is visualised as the origin of the biplot (see Fig. 15, note the small range on both axes of the biplot indicating low degrees of maximum independence on both dimensions 1 and 2).

The intensity of the relationship between 'Semantic domain' and 'Corpus' is very weak (Cramér's $V = 0.08$). This was to be expected because, as shown in previous research (Deshors 2017), the progressive as a construction is associated with broadly similar distributions of semantic domains across registers and varieties. Nevertheless,

CA helps us tease out subtle differences between the profiles of the variables under investigation.

Fig. 15 displays the first two dimensions of the CA results which, together, account for 95.93% of the total inertia of the data. The supplementary variables (which do not contribute to the profiles) are labelled in dark red. The categories plotted furthest away from the axes are the most "different" from the average profiles, in other words, most independent.

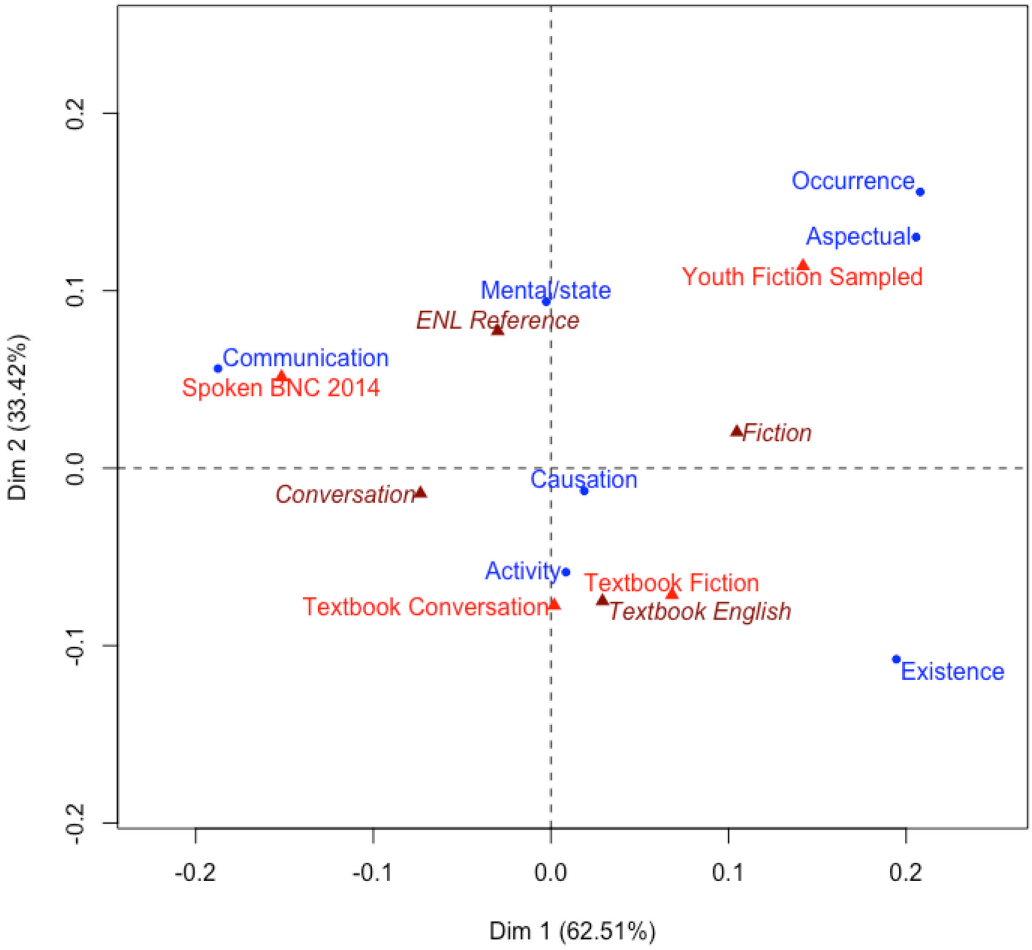


Fig. 15: Simple binary correspondence analysis of 'Semantic domain' and 'Corpus'

Starting with Dimension 1, which explains the most variance, we observe a distinction between the Spoken BNC2014 on the negative end of the scale, and the Youth Fiction corpus on the other. Though following the same register pattern, the textbook registers are much closer to the x-axis and this reflects less pronounced differences in the distribution of semantic domains between the two textbook registers. The semantic category 'communication' clusters with the Spoken BNC2014, while 'occurrence' and 'aspectual cluster' with Youth Fiction.

While Dimension 1 clearly reflects the distinction of register, the spread of corpus points along Dimension 2 shows a distinction between the reference registers (positive

values) and the textbook ones (negative values). The semantic domain 'activity' clearly clusters with both Textbook Conversation and Textbook Fiction, whereas progressives with 'mental' verbs are less associated with Textbook English and more with the reference ENL corpora.

4.3.4.4 Distinctive Collexeme Analyses

All the analyses presented thus far have only considered progressive constructions, comparing the morphosyntactic, functional, lexical and semantic peculiarities of progressives featured in Textbook Conversation and Textbook Fiction as compared to two samples of naturally occurring texts of the same registers. In the following, by contrast, the lexis of progressive constructions will be examined through the lens of the progressive vs. non-progressive alternation by comparing the results of two independent Distinctive Collexeme Analyses (DCAs) for each register under study.

Table 31 presents the results of the DCAs for the comparison Textbook English vs. naturally occurring ENL conversation. In terms of absolute difference, by far the largest difference reported in Table 31 concerns the stative verb BE at the bottom of the table. It is worth noting that the direction of the difference contradicts the findings reported in 4.3.4.4 and Table 32. As a reminder, we observed that, when comparing solely the verbs that occur in the progressive in either textbook or ENL registers, BE appears to be under-represented in both Textbook Conversation and Textbook Fiction compared to their respective reference registers. However, when comparing the number of occurrences of BE as a main verb in FVPs in either progressive or non-progressive constructions (as in Table 31), it transpires that BE is in fact less attracted to the progressive in the Spoken BNC2014 than in Textbook Conversation. This is likely due to the considerably higher frequency of BE as a main verb in FVPs in naturally occurring conversation than in textbook dialogues (66,344 occurrences of BE as a verb pmw in the Spoken BNC2014 vs. 47,399 pmw in Textbook Conversation). The present result thus confirms the indispensability of combining several analysis methods to answer any one research question.

Table 31: Most significantly distinctive collexemes associated with the progressive (positive G^2 values) and the non-progressive constructions (negative G^2 values) in Textbook Conversation and the Spoken BNC2014 sample

Textbook Conversation				Spoken BNC2014 sample			
Lemma	**Non-prog (O:E)**	**Prog (O:E)**	**CL**	**Lemma**	**Non-prog (O:E)**	**Prog (O:E)**	**CL**
WEAR	71:7.2	99:162.8	227.8	TALK	134:14	95:215.9	464.5
TALK	88:12.5	205:280.5	218.0	TRY	127:18	172:281.4	337.1
DO	206:74	1527:1659	177.8	DO	333:182	2766:2917.4	116.1
LOOK	117:44	921:994	90.3	SAY	214:110	1655:1759	87.2
WAIT	50:10	188:228	88.1	GO	325:193	2960:3092	86.6
GO	205:105	2263:2363	83.4	JOKE	15:0.9	0:14	85.1
TRY	55:13	253:295	81.1	LOOK	101:39	556:619	77.3
KID	13:0.7	3:15.3	67.0	WORK	56:18	243:282	59.0
WORK	50:14	279:315	60.2	COME	102:47	695:751	53.6
PLAY	56:18	353:392	58.4	SIT	36:9	113:140	53.4
PLAN	17:2	26:41	52.0	RECORD	15:2	11:5	50.4
JOKE	11:1	5:15	50.1	KID	8:1	0:7.5	45.4
RAIN	12:1	9:20	47.9	PLAN	12:1	9:20	41.0
STAND	20:4	67:83	38.5	RAIN	8:1	2:10	35.1
STUDY	18:3	56:71	36.5	PLAY	30:9	119:140	35.0
LIE	12:1	21:32	34.4	STRUGGLE	9:1	4:13	34.6
WALK	27:8	151:170	32.3	DRINK	18:4	43:58	33.3
SMILE	11:5	21:30.6	30.2	WAIT	22:6	71:88	31.7
PREPARE	8:1	9:16.3	27.8	WEAR	18:4	49:63	30.0
THINK	27:68	1558:1518	-33.1	REMEMBER	1:189	311:294	-29.9
HAVE	59:116	2675:2618	-37.1	GET	121:190	3113:3045	-31.9
LET	3:33	775:750	-47.3	MEAN	2:27	449:425	-40.2
LOVE	0:23	544:521	-47.5	SEE	11:56	943:898	-57.1
SEE	8:58	1350:1300	-70.8	HAVE	94:188	3102:3009	-63.2
WANT	2:50	1176:1128	-86.3	WANT	3:52	873:825	-83.4
BE	16:846	19876:19047	-1962.2	BE	56:1237	21020:19839	-2691.1

The most striking differences between the results of the two progressive vs. non-progressive DCAs are visualised in Fig. 16: each point represents one verb lemma. All verb lemmas are included except BE, which returned very negative CL values in both DCAs (see Table 31). It was therefore removed before the CL values were standardised (to z-scores) to avoid this one verb skewing the results. The scatterplot makes clear that the CL values that emerged from the two DCAs are very highly correlated (Pearson's $r(447) = 0.98,\ p < 0.001$). We may therefore conclude that there are only few verbs that are considerably more or less attracted to the progressive in Textbook Conversation than in the Spoken BNC2014 sample. The labelled points on Fig. 16, however, correspond to the twenty lemmas with the greatest absolute differences in standardised CL values. The rest of the analysis will focus on these verbs.

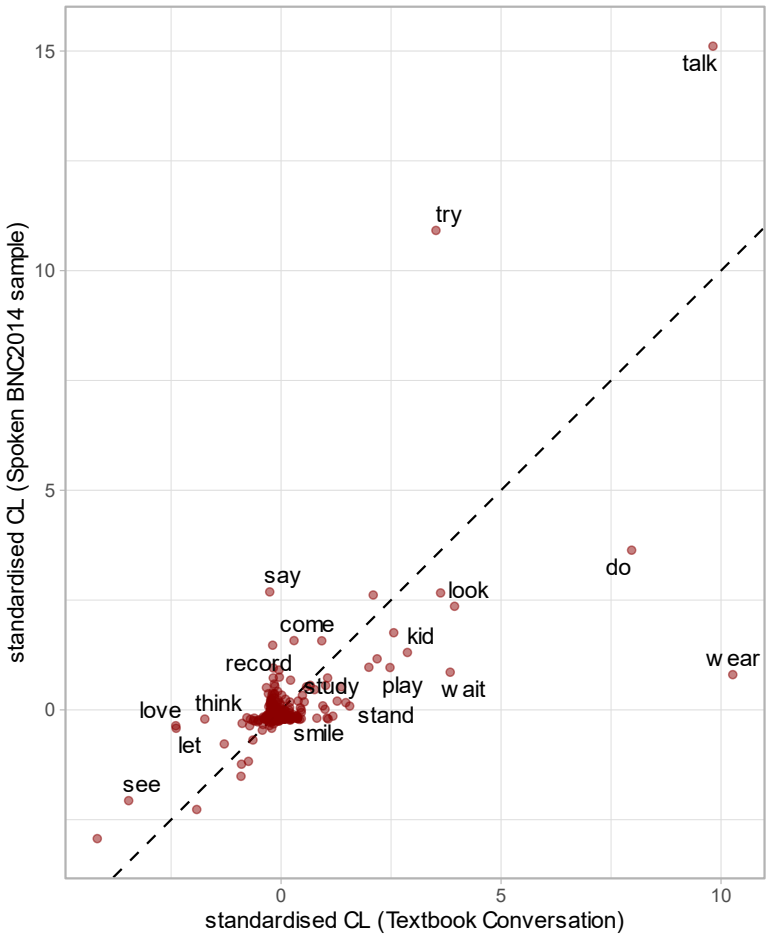


Fig. 16: Comparison of the standardised CL values (scaled and centred G^2) returned by the two progressive vs. non-progressive Distinctive Collexeme Analyses (DCAs). Positive values represent attraction to the progressive, negative to the non-progressive construction.

We will begin by examining the bottom left quadrant with negative standardised G^2 values. These points correspond to verb lemmas repelled by the progressive construction in both corpora. The labelled points in this quadrant are all found to the left of the regression line. These are stative verbs and verbs of involuntary perception that textbooks frequently list as "exception" or "state verbs" that cannot be used in the progressive (see, e.g., Niemeier 2017: 109), e.g., LOVE, LET (in the sense of ALLOW), SEE (as a verb of involuntary sensory perception) and WANT. However, the present results show that, in natural spoken conversation, stative verbs, in particular, do not occur in the progressive quite as rarely as EFL grammars would suggest, e.g., (103)–(106).

(103) so I ended up going and babysitting well the daughter's probably eleven right
something like that erm so I **was** kind of **letting** her be in charge you know <BNC2014: S8BQ>

(104) why don't you let me buy a b Park Lane?
no I'm not letting you buy Park Lane <BNC2014: SXK5>

(105) and yeah she was like oh she she she **was loving** she **was loving** the fact she was being recorded though like <BNC2014: S6BS>

(106) yeah **are** you **wanting** that tape back?
no
are you sure?
yeah you leave it here <BNC2014: S2KP>

Among the verbs that are displayed to the left of the regression line on Fig. 16, we find RECORD, which would suggest that it is considerably more attracted to the progressive in natural conversation than in textbook dialogues. Its presence on this side of the plot has, however, already been identified as a corpus artefact resulting from the data-collection process of the Spoken BNC2014 (see 4.3.4.1 and excerpts (95) and (96)).

The remaining verb lemmas to the left of the regression line on Fig. 16 confirm another finding from the previous analyses: TALK, SAY, THINK and, in many cases, also TRY, are communication verbs, which were already shown to be more frequent in the progressive in natural conversation than in textbook dialogues. Whilst this DCA comparison confirms the findings of the lexical and semantical CovCAs in 4.3.4.1 and 4.3.4.2, as well as those from the CA in 4.3.4.3, it is interesting to note that the rank order of the most mis-represented lexemes is different. SAY is no longer the most under-represented communication verb in the progressive, but rather TALK and TRY.

Among the labelled verb lemmas in the bottom left quadrant of Fig. 16, we also find THINK, which constitutes a special case. It is repelled by the progressive construction in the textbook conversation corpus, but shows no significant association to either the progressive or the non-progressive construction in the Spoken BNC2014 sample. A qualitative exploration of the corresponding concordance lines shows that, in EFL textbooks, THINK is primarily used in non-progressive constructions to give opinions (107)–(108). In addition, textbook dialogues sometimes feature THINK to express doubt (109) or hedge statements (110).

(107) Jack: What about a book? Oliver: No, **I don't think** that's a good idea. <TEC: Access G 1>

(108) I didn't like Celtic music before because **I thought** it was too fast and repetitive. <TEC: Hi There 6e>

(109) Frank Sinatra? Did he write the song?
Hmmm... **I think** so...? I know he sang it. <TEC: Piece of Cake 5e>

(110) Hi, it's me. Where are you? - Right. Listen! I'm at home. Why don't you come over? There's something interesting, **I think**. <TEC: Hi There 6e>

There is no doubt that these functions are also very common in naturally occurring speech. The qualitative concordance analysis revealed that the difference observed in Fig. 16 is, in fact, foremost driven by just one progressive discourse-structuring

phraseme: *I'm/was thinking* (111), which is almost ten times more frequent in the Spoken BNC2014 (1,899.6 occurrences per 10,000 words in the Spoken BNC2014 [n= 2,444]) than in Textbook Conversation (196.7 per 10,000 words [n = 10]).[29]

(111) Stockholm
Copenhagen
Copenhagen yeah I haven't really been to Scandinavia
that's really funny cos I was thinking that as well I was thinking if grandad was here
yeah <BNC2014: S6JP>

Like other phrasemes featuring communication verbs in the progressive, in naturally occurring conversation, *I was thinking* frequently serves to frame discourse (112)–(113).

(112) no I don't think he's in quite that state yet… anyway and then um… what else **what *[is]* he saying**? Oh I said **I was telling** him about John and he was like Oh we should go there on like me you and Matt… we should go… and **I was thinking** um baby?
then **I was saying we were thinking** of going to Sunderland like in September <BNC2014: SR8N>

(113) erm when I went into the love chart I was hoping I might **I was thinking** you might be at the top but then when but you were at the bottom with Jill <BNC2014 SES2>

By contrast, out of the mere eight occurrences found in textbook dialogues, only three have discourse structuring or softening functions similar to those found in the authentic data (114)–(115).

(114) Kim, I've got an idea, but don't laugh. At school today, **I was thinking**: I'm from Seattle but I've never been to top of the Space Needle.
<TEC: Green Line New 4>

(115) Sounds great! So Cape Town, good. Now would you like to go somewhere else? **I was thinking** that we could spend one week in Cape Town and one week at the Kruger National Park in the North East. <TEC: Piece of Cake 4[e]>

Conversely, Fig. 16 shows that another verb belonging to the semantic domain of communication, KID, is more strongly associated with the progressive construction in Textbook Conversation than in the Spoken BNC2014 sample (unlike JOKE, which ranks high on both collexeme lists in Table 31). Whilst textbook authors clearly favour the interrogative phraseme *are you kidding (me)?* and the explosive *you're kidding!,* the Spoken BNC2014 also features the non-progressive phrase *I kid you not.* Before arguing for the inclusion of this particular phraseme in textbook dialogues,

[29] The following Sketch Engine CQL query was used to arrive at these frequencies: [word="I"] [lemma="be"] [lemma!="not"] {0,2} [word="thinking"]. It allows for small variations on the phraseme such as *I'm/am/was just/yeah/er/like/kind of thinking*, but excludes negated forms which do not fulfil the same discourse function.

however, it should be noted that it is very unevenly dispersed across the Spoken BNC2014: 19 out of the 29 occurrences were uttered by just one speaker (S0439, a 23-year-old female L1 speaker).

In a similar vein, the lemma COME is more strongly attracted to the progressive in naturalistic spoken conversation than in Textbook Conversation as a result of two phrasemes frequently featured in textbook dialogues: *come on!* (116) and the phrasebook classic *I come from [country/town]* (117). It also worth noting that the most frequent one-to-the-right collocates of COME in Textbook Conversation are *on*, *to*, *from* and *here*. By contrast, in the Spoken BNC2014, *back*, *out*, *in*, *to* and *up* are most frequently paired with COME and these phrasal verbs are more likely to occur in the progressive, as in (118)–(119).

(116) A rainforest in Cornwall? **Come on**, Lucy, it can't be real... England isn't hot enough for a rainforest. <TEC: Access G 2>

(117) I **come from** South Africa. I'm joyful and friendly. <TEC: Piece of Cake 5e>

(118) if if your doctor occupational health whoever say it needs to be February it'll be February but I'm thinking in my mind **I should be coming back** January <BNC2014: SMW8>

(119) better watch the dog now
are you coming out?
yeah
we'd better put the dogs on a lead cos the chickens are out there
<BNC2014: S2XJ>

To the right of the regression line in Fig. 16 we find verbs that are more strongly associated with the progressive in Textbook Conversation than in the Spoken BNC2014. The case of KID has already been discussed. As noted in the discussion of the results of the previous analyses (see 4.3.4.1), DO, PLAY and STUDY are instantiations of the kinds of activity verbs that are frequently used by textbook authors as model, prototypical verbs for the progressive construction. Similarly, WEAR, STAND and SMILE have been shown to be frequently featured in textbook dialogues modelling picture description tasks designed to provide learners with the opportunity to practise using the progressive.

The remaining lemma on this side of the scatterplot, WAIT, is also associated with how the progressive is traditionally taught in EFL textbooks and grammars: textbook authors make frequent use of WAIT as a means of featuring present and past perfect progressives in dialogues (63). It is also a useful verb to demonstrate the progressive's framing function (120), which was shown in 4.3.3.5 to be over-represented in both textbook Conversation and Fiction.

(120) A full-blown fight broke out between the two groups and someone went to get a teacher. While we **were waiting** out outside the head teacher's

```
office, we got talking and he said something that made me laugh.
<TEC: Solutions Intermediate>
```

In the following, we turn to the results of the same types of analyses as above, now focusing on the Fiction register. Thus, Table 32 compares the results of two progressive vs. non-progressive DCAs for the fiction register.

Table 32: Top 30 verb lemmas with the greatest absolute difference in attraction/repulsion CL scores to the progressive in Textbook Fiction and the Youth Fiction Corpus (negative values in the second and/or third columns represent repulsion of the progressive, in other words, attraction to non-progressive constructions)

Textbook Fiction				Youth Fiction sample			
Lemma	**Non-prog (O:E)**	**Prog (O:E)**	**CL**	**Lemma**	**Non-prog (O:E)**	**Prog (O:E)**	**CL**
SIT	62:12	180:230	120.5	TALK	35:6	85:114	76.4
WEAR	37:4	49:82	112.6	WORK	23:5	73:91	41.2
TALK	50:8	120:162	110.6	JOKE	6:1	0:6	36.4
DO	82:25	429:487	92.3	TRY	36:12	204:229	36.0
WAIT	41:8	130:163	74.1	STAND	31:10	176:197	30.9
STAND	35:8	128:155	56.0	PLAN	8:1	6:14	30.0
WALK	43:14	243:272	43.4	LIE	19:5	77:91	27.5
SHINE	9:1	5:13	36.9	SIT	27:10	167:185	23.9
TRY	34:11	193:216	34.0	FIND	0:13	259:247	-25.7
ASK	3:24	503:482	-31.3	WANT	1:18	372:355	-29.2
WANT	1:28	588:561	-49.8	SEE	7:33	679:653	-31.9
SEE	4:38	794:760	-53.0	HAVE	16:52	1056:1020	-36.0
KNOW	0:29	601:572	-59.8	KNOW	1:33	678:646	-58.7
SAY	24:93	1907:1838	-78.7	SAY	32:122	2506:2416	-104.2
BE	3:348	7227:6883	-775.3	BE	15:334	6919:6600	-644.3

In contrast to the Conversation DCAs (Table 31), the Fiction DCA comparison in Table 32 clearly confirms that Textbook Fiction more strongly associates the verb lemma BE with non-progressive constructions than the Youth Fiction corpus. In fact, BE is the lemma with the greatest absolute difference in CL values. This would suggest that BE as a lexical verb in progressive FVPs appears to be under-represented in Textbook Fiction. This is likely to be problematic for EFL learners since the phraseological pattern *BE + being* is known to be subject to specific lexical restrictions and to fulfil distinct communicative and rhetorical functions. In terms of lexico-grammatical patterning, an earlier corpus study of the BNC1994 revealed that the

vast majority of collexemes in the adjective slot of the *BE + being + adjective* construction (81.21%) are "judgement adjectives" (e.g., *unusual, clever, silly, cautious, irresponsible*) (Kheovichai 2017: 110). Functionally, it would appear that BE + *adjective* is often employed instead of an action verb in order to "assign an evaluative meaning to the action" (Kheovichai 2017: 127). Indeed, a qualitative comparison of random progressive concordance lines with BE from Textbook Fiction and Youth Fiction suggests that, as already noted in 4.3.4.1, this under-representation of BE in the progressive in Textbook Fiction is likely to result in learners being underexposed to the co-occurrences and functional uses of BEING + adjectives describing temporary and/or uncharacteristic behaviours – such constructions being highly frequent in Youth Fiction, e.g., (121)–(122).

(121) He said that **I was being selfish** and that I was never to set foot inside the house again. <Youth Fiction: Haddon 2003: The Curious Incident of the Dog in the Night-Time>

(122) "Well, I don't suppose it matters," sighed Hermione. "Even if **he was being honest**, I never heard such a lot of nonsense in all my life." <Youth Fiction: Rowling 2007: Harry Potter and the Deathly Hallows>

Fig. 17 compares the attraction of verb collexemes to the progressive/non-progressive constructions in Textbook Fiction and Youth Fiction. The circles on or close to the diagonal line behave similarly in both corpora. As in the corresponding figure for the conversation register (Fig. 16), verbs that behave most differently are labelled.

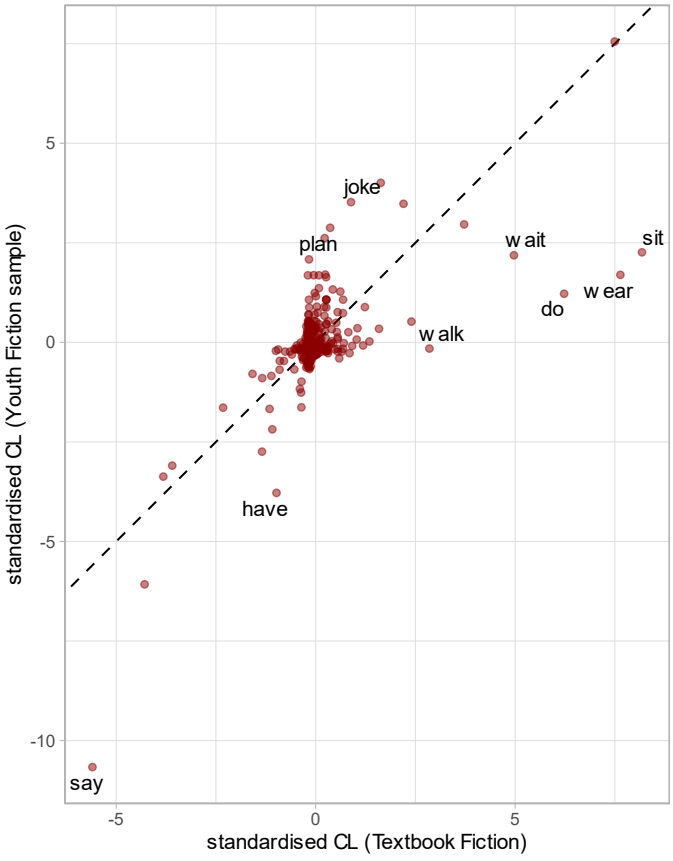


Fig. 17: Comparison of the standardised CL values (scaled and centred G^2) returned by the two progressive vs. non-progressive Distinctive Collexeme Analyses (DCAs)

As was to be expected, 'communication' verbs follow different patterns in fiction than in conversation. For instance, SAY is more strongly associated with non-progressive forms in Youth Fiction than in the corresponding textbook register. This is most likely due to the considerably higher occurrence of the form *said* (7,883 occurrences pmw in Youth Fiction compared to 5,539 in Textbook Fiction). TALK, too, follows the opposite pattern to that observed in the two conversation varieties. In Textbook Fiction, it appears to be over-represented in the progressive compared to Youth Fiction. JOKE, on the other hand, follows the trend observed in Textbook Conversation and is seemingly under-represented in the progressive in Textbook Fiction.

Some of the verbs displayed to the right of the regression line in Fig. 17 are frequently used as a framing verb in Textbook Fiction – a function, which was already shown to be over-represented in Textbook Fiction as compared to Youth Fiction (see 4.3.3.5 and 4.3.4.1). The case of WALK was already discussed in 4.3.4.1 (see also excerpts (100)–(102)). Two additional typical 'framing' verbs to come out of the analysis displayed on Fig. 17 are SHINE (123) and WAIT (124).

(123) It was not a moment too soon. The fog **was rising** and the moon **was shining** on the hills. We heard footsteps and saw a light that went up and down. <TEC: Green Line 3>

(124) When Prince Llewellyn walked through the door, Gelert **was waiting** for him. <TEC: English in Mind 3>

The verbs WEAR and DO follow the same patterns as in all previous analyses – they are more attracted to the progressive in the two textbook registers than in their respective reference corpora. According to Fig. 17, SIT, however, is more attracted to the progressive in Textbook Fiction than in Youth Fiction, although the opposite was observed for the conversation varieties. In 4.3.4.1, it was speculated that WEAR was strongly attracted to the progressive because textbook authors often rely on picture description tasks to practise the progressive and that these tasks are sometimes (usually implicitly) modelled in textbook dialogues. In the case of narrative texts, however, it would appear that the over-representation of WEAR in the progressive is due to textbook authors having to describe the characters of their stories with the limited vocabulary that learners are familiar with. Clothes being one of the first few topics introduced in most textbook series, WEAR becomes the lexical verb of choice for character descriptions. Similarly, DO is a highly versatile action verb which textbook authors can use to describe many activities that the authors included in the Youth Fiction corpus can depict using a much broader range of lexical verbs. Perhaps the same is true of SIT as a simple description verb introduced early on in textbook series. No functional differences in the use of progressive and non-progressive constructions with SIT could be deciphered from a concordance-based analysis.

Finally, one more lemma on Fig. 17 merits further exploration: HAVE is repelled by the progressive construction in both Textbook Fiction and Youth Fiction but the plot shows that it is significantly more so in the narrative texts of the TEC than in the novels of the Youth Fiction corpus. This finding can be traced to the fact that when HAVE is found as the main verb of a verb phrase in Textbook English, it is most frequently in the sense of possessing something, in which sense progressive constructions are, of course, very rare. By contrast, HAVE in novels is most often used as a light verb and in a wide variety of collocations (125)–(126), which can and frequently are used in the progressive (e.g., *having a nap, having a seizure, having an argument, having a chat/conversation, having second thoughts, having a hard time, having trouble*, etc.). Of those many highly frequent HAVE collocations that are often found in the progressive, the only one featured more than once in Textbook Fiction is *having fun* (126), which, alone, accounts for a third of all HAVE progressive constructions.

(125) Harry rather doubted he would be able to bring off this particular spell; **he was still having difficulty** with non-verbal spells, something Snape had been quick to comment on in every D.A.D.A. class. <Youth Fiction: Rowling 2005: Harry Potter and the Half-Blood Prince>

(126) Zoe watched as the dog's enthusiasm quietened. Jake sat on the snow as she snuffled in his ear. It almost seemed to Zoe that the two of them **were having a conversation**. <Youth Fiction: Joyce 2010: The Silent Land>

(127) Back in the Victorian room, Jay **was having fun** in front of the camera. He was a Victorian gentleman, and his in the photo was a fine Victorian lady. <TEC: Green Line New 3>

4.4 Conclusion

In the following, some of the key findings of this chapter are summarised in an attempt to provide answers to the seven research questions laid out in 4.1.3.

The first aim was to verify whether progressive constructions are genuinely overused in Textbook English. The analysis in 4.3.1 showed that this was not the case for either Textbook Conversation or Textbook Fiction texts. As for the second research question, a fine-grained analysis of the morphosyntactic contexts in which EFL learners encounter progressives via their textbooks revealed a number of notable divergences from the reference corpora, particularly in the comparisons between Textbook Conversation and the Spoken BNC2014 sample.

Starting with the morphosyntactic contexts of progressives, past progressive forms were found to be under-represented in Textbook Conversation, even after making allowances for the late introduction of this tense form in school textbook series (see 4.3.2.1). Overall, Textbook Conversation appears to slightly under-represent contracted forms of the progressive. However, the ratio of contracted to non-

contracted progressives varies markedly across different textbook series. Two French textbook series, *Join the Team* and *Piece of Cake*, substantially under-represent contracted progressives (see 4.3.2.2) in their dialogues. In addition, negated progressives were also seen to be under-represented in both Textbook Conversation and Fiction (see 4.3.2.3). In contrast, progressive in interrogative contexts are seemingly over-represented in Textbook Conversation (see 4.3.2.4). Furthermore, the types of questions featured in textbook dialogues and authentic conversations differ quantitatively, both in terms of syntax and rhetorical functions (see 4.3.2.4). Though the proportions of active and passive progressives do not significantly differ across the textbook and ENL registers, the Spoken BNC2014 features more colloquial GET and HAVE progressive passives, whereas Textbook Conversation exclusively features *being* + V_{ed} passive constructions that are more typical of written registers than spontaneous speech (see 4.3.2.5).

The third research question concerned the functions of progressives in the two Textbook English registers. Overall, tenses and time references are distributed similarly across the textbook registers and their corresponding reference corpora when taking the effect of the gradual increase in complexity across textbook levels into account. Still, the results show that progressives in Textbook Conversation less frequently refer to generally valid statements and hypothetical situations than in authentic conversation (see 4.3.3.1). Furthermore, future time references are proportionally more frequently expressed with a present progressive. Modal constructions – foremost WILL + *be* + V_{ing} –, on the other hand, are under-represented compared to the Spoken BNC2014 (see 4.3.3.2). Additional functions of progressive constructions such as referring to repeated actions, non-continuous actions and gradual change or development are also under-represented in both textbook registers (see 4.3.3.3–4.3.3.5). Progressives used in framing contexts, however, are considerably over-represented in textbook registers, in particular in Textbook Conversation as this function was shown to be very rare in the transcripts of the Spoken BNC2014 (see 4.3.3.5).

The fourth research question focused on Textbook Conversation, asking whether the progressives featured in the dialogues of modern textbooks are now more representative of present-day authentic ENL conversation than was the case at the time of Römer's (2005) study. Whilst some of the trends were confirmed, others were not. For instance, whereas Römer (2005: 244) reported that the present progressive, was overused in her corpus of textbook dialogues, 4.3.2.1 made clear that this trend is only true of the beginner textbooks. Furthermore, it was suggested that there is good pedagogical reasoning behind this choice because present progressives are by far the most frequent form of progressives in Spoken English. Similarly, perfect progressives were not found to be under-represented in the present Textbook Conversation subcorpus. However, past progressives are still under-represented across

all levels of textbook dialogues. Overall, contracted progressives seem to no longer be as under-represented as they were at the time of Römer's study; however, interestingly, the two most recent versions of the German textbook series *Green Line* (pub. 2006-2009 and 2014-2018) still appear to under-represent contracted progressives (see Table 17), though noticeably less so than in the older version examined by Römer (pub. 1995-2000, see Römer 2005: 245). Likewise, though negated progressives remain marginally less frequent in textbook dialogues than in naturally occurring British English conversation (see 4.3.2.3), the difference is not as striking as in Römer's data (2005: 250). Contrary to expectations, the more recent series from Klett, *Green Line New*, features the second lowest proportion of negated progressives (3.65%) in its dialogues, whereas the older version, *Green Line*, has the highest proportion, 7.19%, which is remarkably close to that observed in the Spoken BNC2014 reference corpora. As in Römer (2005: 252), progressives in interrogative contexts were found to be over-represented in textbook dialogues.

Turning to the function of progressives, Römer (2005: 262) observed that progressives that express repeatedness were comparatively rare in her textbook dialogues (9.12% and 9.87%, compared to 35.05% and 38.64% in her reference ENL data) and argued that, since she identified repeatedness as one of the core functions of the progressive, it should be more frequently and prominently featured in Textbook English (Römer 2005: 284). The results presented in 4.3.3.4 (see Table 24) appear to show that, whilst the repeated function of progressives is now better represented in modern textbooks, it remains underused. Although 'non-continuous' progressives were even rarer in the older textbook dialogues (2.53% and 4.56%), for this functional feature, too, we can confirm the tendency observed by Römer (2005: 261) (see 4.3.3.3). Finally, like Römer (2005: 267), we still observe a strong tendency to emphasise the framing function of the progressive in textbook dialogues, in spite of the fact that it is quite atypical of spoken English (see 4.3.3.5).

To answer the fifth research question, three types of collostructional analyses were conducted and compared. The ensuing results and discussion thereof demonstrated the benefit of using a range of methodological approaches to tease out relevant patterns and avoid biases (see Deshors 2017 for a similar plaidoyer). Verbs such as WEAR and LOOK were found to be consistently over-represented in the progressive, as were verbs associated with the framing function, e.g., WALK and WAIT. Unsurprisingly, stative verbs and other verbs typically taught as not (usually) occurring in the progressive are more strongly associated with non-progressive constructions in the textbook registers than in the respective reference corpora (e.g., BE, LIKE, LOVE).

The CovCAs and CA conducted in an attempt to answer the sixth research question revealed some interesting patterns as to the semantic domains of the verbs featured in the progressive in the two textbook registers. Although the distributions of

semantic domains was broadly similar across all four (sub)corpora under study (see 4.3.4.1), subtle differences of pedagogical relevance were identified. For instance, 'activity' verbs are more associated with textbook registers, whereas 'mental/state' verbs are more associated with the ENL reference corpora. The most striking divergence, however, was observed with the semantic domain of 'communication' in the Conversation register (see 4.3.4.3). Indeed, verbs such ARGUE, JOKE, TALK, TRY and SAY were consistently found to be less strongly associated with the progressive in Textbook Conversation across all analyses.

Finally, we asked whether the differences noted in the use of progressives across Textbook English and the ENL reference corpora made pedagogical sense, or whether they might result in learners being unfamiliar with frequent, and thus arguably useful, progressive constructions. Throughout the discussion of the results, it was made clear that many of the differences, especially in terms of tense form and time reference distributions, are probably pedagogically well-founded. Similarly, it may be argued that since question syntax poses a lot of problems for EFL learners, an over-representation of this form in the progressive might also serve a sound pedagogical purpose. In contrast, the notable under-representation of many typical discourse phrasemes featuring progressives seems more problematic. It is likely to result in learners sounding unidiomatic and, for some, may even contribute to their struggle to engage in natural, spontaneous conversation with proficient speakers of English. Similarly, learners' lack of exposure to central functions of the progressive such as repeatedness or gradual change may lead to genuine difficulties in listening and reading comprehension. Some of the differences observed in the lexical verbs associated with the progressive appear to stem from textbook authors' avoidance of many idiomatic light verb constructions (e.g., HAVE *an* argument/chat/*conversation*), which frequently call for the progressive in both conversation and fictional writing. The next chapter homes in on textbook authors' apparent (see 2.2.1.2) underuse of idiomatic collocations, including light verb constructions and phrasal verbs, in secondary school EFL textbooks as it examines the use of MAKE in Textbook Conversation and Textbook Fiction.

5 Making sense of MAKE in Textbook English

Hey look, some words with real meaning on a shopping street! The sign
makes you stop and think, doesn't it? It reads like a poem.
Who needs real meaning on a shopping street?! I'm more into songs anyway.
<TEC: Green Line New 3>

This case-study chapter[30] focuses on the high-frequency verb MAKE, exploring its semantic and phraseological representation in the Conversation and Fiction subcorpora of the TEC as compared to the Spoken BNC2014 and the Youth Fiction corpora. Presentations of MAKE in Textbook English are additionally compared to the results of relevant learner English corpus studies.

5.1 Rationale and aims

MAKE is typically one of the first verbs to be taught in EFL lexical syllabi. Its monosyllabic form and (often mistakenly) assumed equivalence with equally frequent verbs in other Indo-European languages such as *faire*, *machen* and *hacer*, has contributed to MAKE being commonly perceived as an "easy verb" for English learners to master (see, e.g., Lennon 1996; Gouverneur 2008a: 223; Liu & Shaw 2001: 188). However, it is well known that with high frequency comes polysemy and, in the case of MAKE, this involves delexical, phrasal, prepositional and other idiomatic uses. In addition, just like *faire, machen* and *hacer,* MAKE enters into many restricted collocations (as defined by Cowie 1998) that often cannot be translated by simple one-to-one equivalences (e.g., Dirven & Radden 1977), hence generating a high potential for L1 interference in L2 learners' production, as illustrated in (128)–(131) from the French and Spanish L1 subcorpora of the Open Cambridge Learner Corpus (Cambridge University Press 2017, available on Sketch Engine).

(128) It has become impossible nowadays to **make some research** [~ *faire des recherches*] without the Internet. <OCLP: L1_FR, 359, C2>

[30] Preliminary results of the analyses from this chapter were presented at the IVACS 2019 symposium in Dortmund. Parts of this chapter have been published as: Le Foll, Elen (2022). *MAKING tea and mistakes*: The Functions of MAKE in Spoken English and Textbook Dialogues. In Zihan Yin & Elaine Vine (eds.), *Multifunctionality in English: Corpora, Language and Academic Literacy Pedagogy*, 157–178. Oxon; New York: Routledge and Le Foll, Elen (2023). Textbooks and Corpus Linguistics: the case of causative constructions. In Kieran Harrington & Patricia Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*. Palgrave MacMillan.

(129) Step by step, you decide to work all of the night, and at the end you forget your body and begin to die. It **has no sense** [~ *avoir du sens* → make sense]; you must respect your own life. <OCLC: 5411, B1>

(130) I have been waiting for the T.V. repair main [sic] on Wenesday A.M. but he didn't come at all. I phoned the firm and they have **took a new appointment** [~ *prendre un rendez-vous* → make an appointment] in two weeks time. <OCLC: L1_FR, 4440, B1>

(131) It was dark, the weather was quite bad, the wind was hitting the windows and I was thinking on what day I had choosen to **make a party** [~ *hacer una fiesta*]. <OCLC: L1_FR, 11372, B1>

(132) Sometimes I forced myself to **make some work** [~ *hacer algunos trabajos*] I had left behind and suddenly I found myself as a selfish mother that cares more about work than about her son. <OCLC: L1_ES, 7664, C2>

Moreover, MAKE is often perceived as the most prototypical causative verb (Altenberg 2002: 99). Bearing this in mind, Gilquin (2016a: 236) hypothesises that MAKE may "serve as a pathbreaking verb in instruction, that is, a verb that is used to introduce the general characteristics of the construction, before the construction is extended to other verbs". Indeed, causative constructions with MAKE (e.g., *it makes you stop and think*) are both highly frequent and of high communicative value; however, they also form complex syntactic patterns which often prove problematic for learners of English (e.g., Altenberg & Granger 2001; Gilquin 2012; 2016a; Liu & Shaw 2001; Wong 1983). Such syntactic errors have long been recognised in mainstream English Language Teaching (ELT) literature (e.g., error note on MAKE in the *Longman Dictionary of Common Errors*, Turton & Heaton 1996, cited in Gilquin 2012: 42). These various potential pitfalls have inspired a plethora of corpus-based studies on English learners' unidiomatic use of MAKE (e.g., Altenberg & Granger 2001; Gilquin 2016a; Liu & Shaw 2001; Nesselhauf 2004).

A selection of the many studies exploring the use of MAKE in learner Englishes is surveyed in 5.1.1. In contrast to this wealth of learner English studies, only one study (Gouverneur 2008b) has, to the present author's best knowledge thus far, delved into the representation of MAKE in what is arguably EFL learners' first and main source of language input: their textbooks. Key findings of this precursor to the present study are summarised in 5.1.2. In light of those conclusions, as well as the findings of past learner English studies (see 5.1.1), the present chapter seeks to explore the following research questions:

1. Is the verb MAKE significantly over- or underrepresented in the conversation and fiction texts of contemporary secondary school EFL textbooks as compared to target language reference corpora of similar registers?
2. Which semantic functions does MAKE in Textbook English typically fulfil? Are these functions representative of the most common meanings of MAKE?

3. Which collocates and semantic fields are typically associated with MAKE in textbook conversation and fiction texts? How does this compare the language of similar registers that learners are likely to encounter outside the EFL classroom?
4. To what extent are known L2 English learners' unidiomatic usage, underuse or overuse of the verb MAKE (see 5.1.1) potentially traceable to a lack of adequate linguistic input in EFL textbooks?
5. Which of the differences noted in the representations of MAKE across Textbook English and the target language reference corpora may be pedagogically well-founded? Are any susceptible of depriving learners of valuable exposure to frequent forms with particularly useful communicative functions?

The methodology to do so is outlined in 5.2. The results of this comparative, Textbook English vs. target 'real-life' language, are analysed and discussed in 5.3. In the interest of clarity, this section is subdivided according to the broad semantic categories identified and compared in 5.3.2: MAKE in the sense of 'produce' (5.3.3), as a delexical verb (5.3.4) and as part of phrasal verbs[31] (5.3.5). Section 5.3.6 examines the syntactic and phraseological patterns of causative MAKE constructions in EFL textbooks. The chapter concludes with an outline of the key pedagogical implications of the present findings (5.4).

5.1.1 MAKE in Learner English

Altenberg and Granger (2001: 176–178) report that French learners of English underuse MAKE in their essay writing and, in particular, significantly underuse delexical and causative uses of the verb as compared to ENL students. Further, both French and Swedish learners appear to underuse MAKE collocations with 'speech' or 'verbal communication' noun objects (e.g., *argument*, *claim, point, statement*).

Nesselhauf (2004) focused on German L1 speakers' use of light verb constructions (referred to as 'delexical' verbs in this chapter, see 5.2) with the verbs HAVE, TAKE, GIVE and MAKE. Since they are based on a relatively small learner corpus (a preliminary version of ICLE-German subcorpus with some 150,000 words; Granger et al. 2009), the results must be interpreted with some caution; however, they nevertheless suggest that, out of these four high frequency verbs, German learners are most liable to making errors in support verb constructions featuring MAKE.

In addition, a number of studies (Altenberg & Granger 2001; Gilquin 2012; 2016a; Liu & Shaw 2001; Wong 1983) have pointed to English learners' tendency to frequently make infelicitous syntactic and lexical choices when producing causative constructions with MAKE. For instance, Gilquin (2016a) compared the lexical choices of ENL, ESL and EFL writers for the non-finite verb slot in causative constructions.

[31] As is most frequently the case in ELT contexts, the term 'phrasal verbs' here refers to both phrasal and prepositional verbs. Thus, the term encompasses *verb + adverb* (*MAKE off*), *verb + preposition* (*MAKE for*), and *verb + adverb and preposition* (*MAKE away with*) constructions.

Using distinctive collexeme analysis (see 4.2.3.2), she was able to identify and quantify the lexical idiosyncrasies observed in MAKE causatives in ESL and EFL student writing. One important finding is that, whilst both types of learners overuse a small number of verbs (foremost, the statives BE and BECOME) and underuse others (in particular, SEEM and APPEAR) in the non-finite verb slot of [X MAKE Y V_{inf}] constructions, EFL learners overuse a number of other verbs compared to ENL students, including BELIEVE, GO AROUND and FEEL (Gilquin 2016a: 243). Gilquin (2012; 2016a) identifies three possible causes for learners' unidiomatic use of periphrastic causative constructions: a) a lack of register awareness, b) negative L1 transfer and c) the inadequacy of teaching materials. This final potential factor, also identified by Altenberg & Granger (2001: 184) in the context of EFL learners and causative constructions with MAKE, will be at the heart of this case study on the representations of MAKE in school EFL textbooks.

5.1.2 MAKE in Textbook English

As mentioned in the introduction to this chapter, Gouverneur (2008b) investigated the selection and presentation of the phraseological patterns of MAKE (and TAKE) in Textbook English. The analysis was based on a subset of the TeMa corpus (Meunier & Gouverneur 2009), consisting of six intermediate and advanced English for general purposes (EGP) textbooks. It focused exclusively on the exercises included in these textbooks and their accompanying materials, i.e., the student workbooks. The detailed pedagogical annotation of the vocabulary exercises in the TeMa corpus singles it out from other textbook corpora: over 80 tags were manually applied to the exercises to encode the type of learning activity the learners engage in (e.g., match words and definitions, complete sentences) and the way each target lexical item is to be completed within the exercises (e.g., select from a word box above task, cross out incorrect answer within the text, retrieve from the mental lexicon) (Meunier & Gouverneur 2009: 189–195). From this vocabulary exercise subcorpus, all 298 occurrences of MAKE were automatically retrieved and manually coded for meaning and phraseological patterning. Gouverneur's fine-grained analysis then focuses on verb-noun restricted collocations with MAKE. Phrasal verbs were not included in this study.

According to Gouverneur's classifications (2008b: 233), by far the most frequent use of MAKE in the TeMa exercise subcorpus is in restricted collocations, which, in her study, included delexical uses and the senses 'do/perform' and 'earn'. At first, this finding seems promising, given that such collocations are known to be considerably more problematic for learners of English than MAKE in its prototypical sense of 'produce' (e.g., Altenberg & Granger 2001: 189). However, whilst intermediate textbooks often place direct emphasis on these phraseological patterns, such explicit treatment of these collocations is much rarer in the advanced textbooks of the TeMa corpus (2008b: 234–235). Furthermore, intermediate-level exercises more frequently

focus on the verb slot in these constructions than the tasks from the advanced textbooks, which tend to focus on the collocate object slot. Learner corpus studies, however, have shown that even advanced L2 users more often struggle with the choice of verb (e.g., Nesselhauf 2004: 71) than with the choice of object noun. Finally, very few of the phraseological patterns with MAKE observed in the TeMa subcorpus were common to all textbooks of the same level (15% across the three intermediate textbooks and 7% across the advanced textbooks), pointing to a lack of systematicity in materials authors' choices.

5.2 Methodology

This section describes the methodology designed to compare the representations of MAKE in Textbook Conversation and Textbook Fiction with naturally occurring ENL data. First, all instances of MAKE as a verb were retrieved from the Fiction and Conversation subcorpora of the TEC. In total, 674 occurrences of MAKE were extracted from the Conversation subcorpus, and 392 from the smaller Textbook Fiction subcorpus. In addition, the same number of occurrences were randomly sampled from the Spoken BNC2014 ($n = 674$), as well as from the Youth Fiction Corpus ($n = 392$). Subsequently, the concordance lines corresponding to these 2,132 occurrences of the verb MAKE were coded for a) meaning, b) main collocate lexemes, and c) semantic field attributed to the collocates.

In a preliminary round of coding for meaning, the semantic categories proposed by Altenberg & Granger (2001) and Gouverneur (2008b) (themselves based on well-known learner dictionaries) were adopted. However, these categorisation schemes proved to be rather susceptible to subjective judgements and there was too little information in these publications to follow exactly the same coding procedure and thus obtain genuinely comparable data. Having compared alternatives, it was ultimately decided to apply the meaning categories derived from the *Valency Dictionary of English* (hereafter VDE) (Herbst, Heath & Roe 2013: 513–517) whose categories are more thoroughly delimited and include exemplifications of the most frequent phraseological patterns (see Table 33). Nonetheless, in some cases, the distinction between the first two categories of the VDE's MAKE entry, 'Ai produce' and 'Aii delexical' (see Table 33), remained problematic. Indeed, over the course of the coding process, it became clear that the term 'delexical', though highly frequent in both the English linguistics and ELT literature, is rarely defined; linguists seemingly operationalise the term in a myriad of ways. In an attempt to remedy this, the present study adopted the definition and examples provided in the *Collins Cobuild English Grammar* (Sinclair et al. 1990: 147–151):

> [Delexical verbs are] very common verbs which are used with nouns as their object to indicate simply that someone performs an action, not that someone affects or creates something. These words have very little meaning when they are used in this way […] In many cases, there is a verb which has a similar meaning to the meaning of the delexical structure. […]

> When you use the word as a noun in a delexical structure, you are naming an event, something which is complete. This structure often seems to be preferred to a structure in which the verb has greater prominence. Note that the verb which corresponds to the delexical structure is often intransitive (Sinclair et al. 1990: 147).

Table 33: Semantic categories of MAKE according to the VDE (Herbst, Heath & Roe 2013) illustrated with sentences from the TEC

Semantic category	Meaning	Example sentence
Ai: Produce	Construct or create	I want to **make a carrot cake**. <TEC: Solutions Elementary>
Aii: Delexical produce	Performing an action	People always **make the mistake** of thinking that […] <TEC: New Missions 2[de]>
B: Earn	Money	But I've **made** enough **money** to buy them a house […]. <TEC: Green Line 2>
C: Achieve	Typically used with destinations or targets to be reached	There's a lot of pressure and stress, but **I've made it** […] <TEC: Green Line New 5>
D: Be/become	Particularly suited for a role or function	I think women **make better poets** than men. <TEC: Solutions Intermediate>
E: Cause	Cause	[T]hey don't **make me wear** things that I don't like. <TEC: Join the Team 3[e]>
F: Ensure	MAKE sure/certain	So the teacher can **make sure** the student really understands. <TEC: Solutions Intermediate>
I: Idioms	Other idiomatic phrases	Maybe I can **make new friends** there. <TEC: Access G 1>
P: Phrasal verbs	MAKE followed by one or more prepositions/particles	[…] I always thought they **made that up** to scare us. <TEC: Green Line 5>

In the manual coding scheme of the MAKE concordance lines, the variable 'collocate lexeme' was designed to capture the object, most frequently a noun, in MAKE occurrences of the semantic categories A-D (e.g., *cake, mistake, money, it* and *poets* for the example sentences listed in Table 33). It was also used to record the non-finite verbs in the periphrastic causative constructions of category E (e.g., *wear*) and the relevant prepositions/particles for category P (e.g., *up*). The collocate lexemes were automatically lemmatised using the procedure already described in 4.2.2.1.

In addition, the semantic fields attributed to the 'produce' collocates were initially automatically tagged with the USAS Wmatrix3 English web tagger (Rayson 2018). To begin with, only the 21 upper-level major discourse fields of the tagset were considered (Archer, Wilson & Rayson 2002). These category codes were manually verified, corrected and refined. To better capture the variance of the data, the following adjustments were made to the original Xmatrix classification scheme: owing to the large number of MAKE collocates belonging to category F ('Food'), this category was subdivided according to the tagset's subcategories F1 ('Food') and F2 ('Drink').

Similarly, category B ('the body and the individual') was subdivided into 'Body' (e.g., *fingerprint, hair*) and 'Clothing & Accessories' (e.g., *bracelet, costume*). Subsequently, the categories 'Body', L ('Life and living things') and W ('World and environment') were combined to form a new 'World, Life & Body' category. Due to very low figures and partially overlapping terms, categories I ('Money and commerce in industry') and Y ('Science and technology') were merged into one 'Industry & Technology' category, whilst E ('Emotion'), G ('Government and public'), N ('Numbers and measurement') and P ('Education') were merged into a 'General & abstract terms' category.

The following section (5.3) presents the results of the analyses of the use of MAKE in Textbook English conducted on the basis of the 2,132 annotated concordance lines. It provides an overview of the frequencies of MAKE across all the register subcorpora of the TEC (5.3.1), before focusing on MAKE in the Conversation and Fiction subcorpora. The final results subsection, 5.3.6, explores the syntactic and phraseological patterns of MAKE as a causative verb across all textbook registers.

5.3 Results and discussion

5.3.1 MAKE in the Textbook English Corpus (TEC)

As a verb, MAKE is typically first introduced within the very first pages of beginner textbooks. Across all textbook series and countries of use of the TEC, learners first encounter the verb MAKE in the context of instructions, e.g., (133)–(139). Correspondingly, the strongest object collocates of MAKE in the beginner textbooks in the TEC (level A, see Table 3) are *list, sentence, note* and *dialogue*. Thus, we see that these early exposures to this multifunctional verb already include delexical uses of MAKE, e.g., (138)–(139).

(133) **Make groups** of three. <TEC: Access G 1>

(134) Practise: **Make sentences** to describe your house. <TEC: Piece of Cake 6e>

(135) Put the words in the correct order and add do or does to **make questions**. <TEC: Achievers A1>

(136) Use the pictures to **make conversations**. <TEC: English in Mind Starter>

(137) **Make notes** while you listen. <TEC: Green Line New 1>

(138) **Make a list** of the characters. **Make a list** of Tom's mother's different problems. <TEC: Join the Team 6e>

(139) Look at the posters and **make suggestions**. <TEC: Hi There 6e>

Across the entire Textbook English Corpus, MAKE is the 9th most frequent verb lemma after BE, DO, HAVE, GO, GET, SAY, USE, and THINK. Moreover, it is the 45th most frequent lemma across all parts-of-speech with a relative frequency of occurrence

across all textbook registers of 1,407 pmw[32]. Per 10,000 verbs, MAKE has a relative frequency of 144.2 (SD = 56.81). That said, Table 34 shows that its distribution across the different registers of the TEC is heavily skewed towards instructional language, which, it is worth recalling, accounts for just over a fifth of the total English word count in the TEC (see Fig. 2).

Table 34: Frequency of the verb lemma MAKE in the TEC

TEC subcorpus	MAKE (n)	MAKE (per million words)	MAKE (per 10,000 verbs)
Conversation	699	1,368	87
Fiction	395	1,556	93
Individual words/sentences	1,776	1,943	128
Informative texts	713	2,354	162
Instructional texts	2,310	3,902	259
Other texts	29	2,015	140
Personal correspondence	115	1,714	103
Poetry & rhyme	82	3,131	182

The following fine-grained analyses will home in on the two textbook registers with the lowest relative frequencies of MAKE, namely Fiction and Conversation. As laid out in 5.2, the representations of MAKE in these two textbook registers are compared to the Youth Fiction corpus and the Spoken BNC2014 respectively.

The first comparison concerns the relative frequencies of MAKE in these two textbook registers as compared to their corresponding ENL reference corpora. The results of this analysis are displayed in Fig. 18. It provides clear evidence that, across both language varieties, the verb MAKE is more frequent in fiction than conversation. However, we note that the relative frequency of MAKE is considerably lower in Textbook Fiction. By contrast, MAKE is more frequent in Textbook Conversation than it appears to be in naturally occurring ENL conversation. Indeed, Fig. 18 shows that the difference in relative frequencies is considerably less pronounced between the textbook registers than it is between each ENL register, a finding that echoes that observed with progressives (see 4.4). Chi-square tests (with Yates' continuity correction) applied to the observed and expected raw frequencies indicate that these differences between each textbook register and their corresponding reference corpora are significant; however, the effect sizes are minimal, suggesting that these differences are unlikely to be pedagogically relevant (Conversation: $\chi^2(1) = 20.22$, $p < 0.001$, $\varphi = 0.003$; Fiction: $\chi^2(1) = 17.26$, $p < 0.001$, $\varphi = 0.003$).

[32] Calculated using Sketch Engine by searching for the lemma MAKE with a verb POS in the TEC.

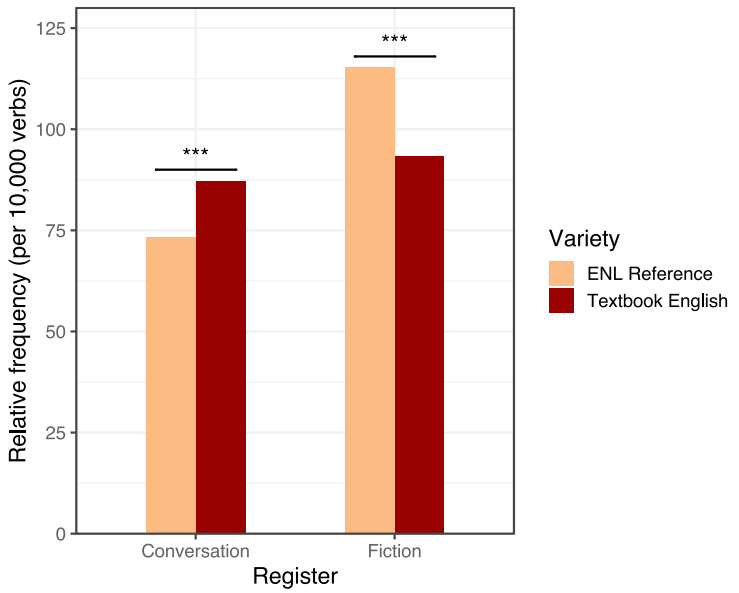


Fig. 18: Relative frequencies of the verb lemma MAKE in Textbook and Reference subcorpora

5.3.2 Semantics of MAKE in Textbook Conversation and Fiction

High-frequency verbs such as MAKE are not only highly polysemous, but they also enter into a vast range of (semi-)restricted lexico-grammatical constructions. This section explores the quantitative differences in the proportional use of these different meanings and constructions in the same two Textbook English registers as compared to similar registers in naturally occurring ENL speech and writing.

Fig. 19 visualises the different proportions of meanings of MAKE in Textbook Conversation and in the Spoken BNC2014 sample. Statistical differences between the proportions of each category were tested using Fisher's exact test for count data. On all the plots featured in this chapter, the resulting p-values are plotted using the following key: $p < 0.001$ (***), 0.001–0.01 (**), 0.01–0.05 (*), 0.05–0.1 (˘).

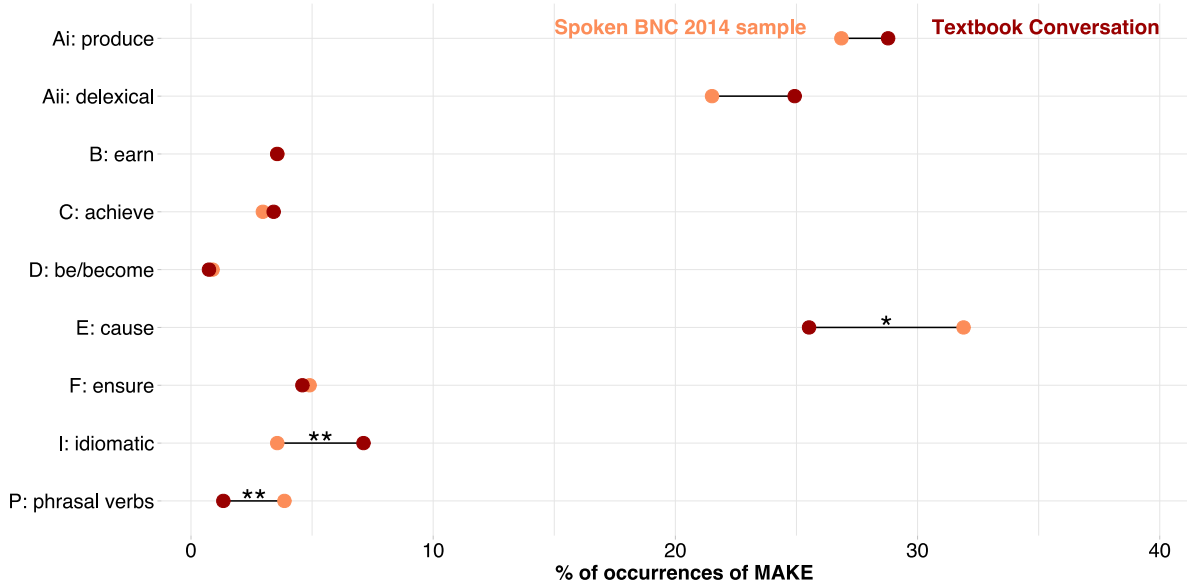


Fig. 19: Comparison of the distribution of MAKE meanings in Textbook Conversation and the Spoken BNC2014 sample

Causative uses of MAKE represent the most frequent semantic category in the Spoken BNC2014 sample. The proportion of 'causative' MAKE occurrences in Textbook Conversation, however, is significantly lower (p = .01, OR = 1.36, 95% CI 1.07–1.74). Causative uses of MAKE will be further explored in 0. The next two most frequent categories, the 'produce' sense and delexical uses, are similarly distributed across the two conversation corpora. The significantly higher proportion of idioms featuring MAKE in Textbook Conversation ($p = .004$, OR = 0.48, 95% CI 0.28–0.81) is primarily due to the very high frequency of the idiomatic phrase *to MAKE friends* (see Table 35). This is clearly a strong topical focus in many of the teenage dialogues written for school textbooks, e.g., (140)–(141).

Table 35: Distribution of the most frequent idioms featuring MAKE in Textbook Conversation and the Spoken BNC2014 sample (absolute frequencies)

Idiomatic phrase	Textbook Conversation	Spoken BNC2014 sample
MAKE friends	25	5
MAKE the best/most of sth.	4	2
MAKE faces	4	4
MAKE fun of sth./sb.	3	1
Practice makes perfect.	3	0
MAKE up YOUR mind	5	3

(140) Dave: That's nice for people on holiday - but I'll be in a new school, and there'll be nobody I know. It'll be horrible. And I'm sure Sid will hate it too.
Jay: Don't worry, you'll **make lots of new friends**. <TEC: Green Line New 2>

(141) Finally, although school trips are fun and help students to relax and **make friends**, sport is even better for doing these things. All in all, I take the view that schools should spend more on sport than on music or school trips. <TEC: Solutions Pre-intermediate>

Fig. 20 compares the distributions of MAKE meanings in Textbook Fiction and a sample of the Youth Fiction Corpus. The most striking difference is found in the significant underrepresentation of delexical uses of make in Textbook Fiction as compared to the Reference Youth Fiction corpus (p = .04, OR = 1.42, 95% CI 1.01–2.01). A more fine-grained analysis of delexical uses of MAKE in Textbook English will be presented in 5.3.4.

Moreover, both textbook registers under study, Conversation and Fiction, feature significantly fewer phrasal verbs with MAKE than their corresponding ENL reference corpora (see Fig. 19 and Fig. 20). Since phrasal verbs are known to represent a

stumbling block for many learners of English as an L2, this is pedagogically relevant finding that will be further elaborated on in 5.3.5.

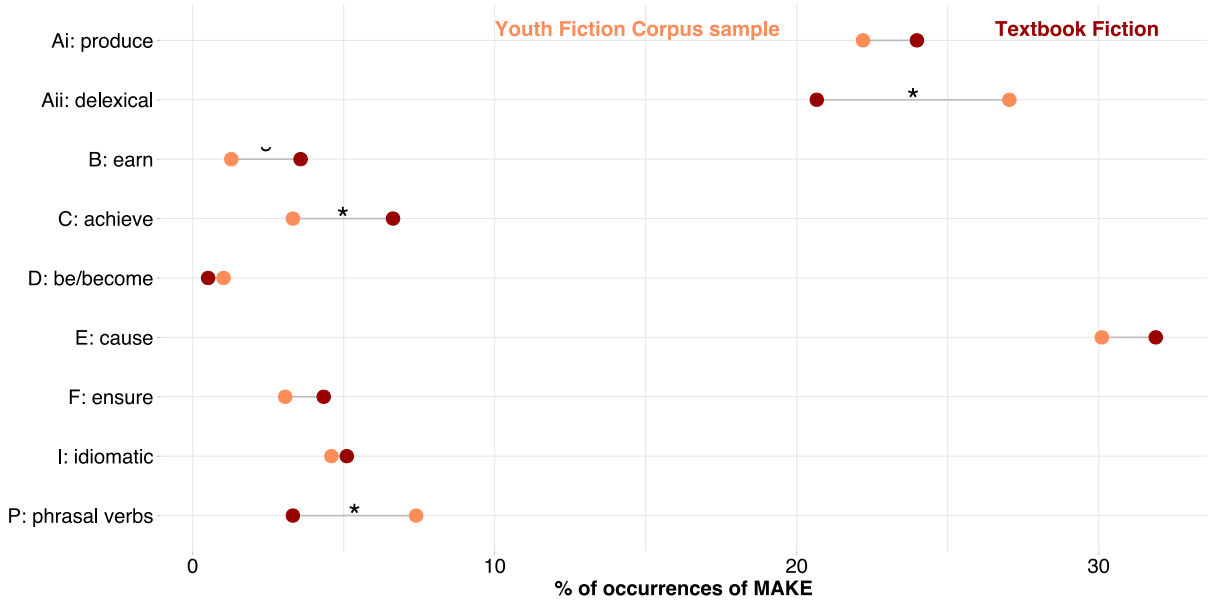


Fig. 20: Comparison of the distribution of MAKE meanings in Textbook Fiction and Youth Fiction

Quantitatively, the semantic category 'achieve' appears to be overrepresented in Textbook Fiction as compared to the Youth Fiction reference corpus. This is primarily due to the fact that, in Textbook English, the 'achieve' category is largely dominated by the phrase *to* MAKE *it*, e.g., (142)–(143).

(142) And we did have enemies - both the white soldiers and the Omaha warriors, who were always trying to capture Kaw boys and girls undergoing their endurance test. It was an exciting time." "What happened **if you couldn't make it**?" Roger asked. <TEC: Green Line 4>

(143) "**They'll never make it**," he whispered. "Teresa has a bad leg and the kids... they can't outrun those things. <TEC: Join the Team 3e>

This phrase is found to occur 17 times in Textbook Fiction, as opposed to just three times in the Youth Fiction sample. By contrast, a qualitative analysis of the corresponding concordance lines reveals that, in the Youth Fiction sample, the 'achieve' category includes a much broader range of intransitive motion constructions (see Goldberg 1995: 3; Rohde 2001; Stefanowitsch 2013), e.g., (144)–(146). These constructions, typical of narrative prose, and syntactically relatively complex because they usually contain one or more adverbs or prepositions, are, however, conspicuously absent from the Textbook Fiction subcorpus. Extract (146) exemplifies the *way*-construction (see Goldberg 1995: chap. 9; 1996; Israel 1996; Traugott & Trousdale 2013), in which MAKE has a "privileged status" as its strongest collexeme, accounting for some 20% of all occurrences of the construction, and as the first verb used in the *way*-construction before it was extended to other lexical verbs (Goldberg 1996: 39). This construction is largely absent from Textbook Fiction.

(144) […] Jack came out of the barn and **made straight for me**.
<Youth Fiction: Douglas 1979: The Hitch Hikers Guide to Galaxy>

(145) She and the bird started to **make off towards my ship**.
<Youth Fiction: Lewis 1952: The Chronicles of Narnia: The Voyage of the Dawn Treader>

(146) Then he **made his way down the stairs** and into the locker room.
<Youth Fiction: Blume 1974: Blobber>

5.3.3 MAKE in the prototypical 'produce' sense

Since the prototypical 'produce' sense of MAKE (sense Ai in the VDE) accounts for a considerable proportion of the occurrences of MAKE in Textbook English, this section delves further into the semantic fields to which the object noun collocates of MAKE in this primary 'produce' sense can be attributed. Fig. 21 first compares the distributions of the semantic fields attributed to the noun collocates of MAKE in the 'produce' sense in Textbook Conversation and the Spoken BNC2014. The semantic categories are ordered from least to most similarly distributed.

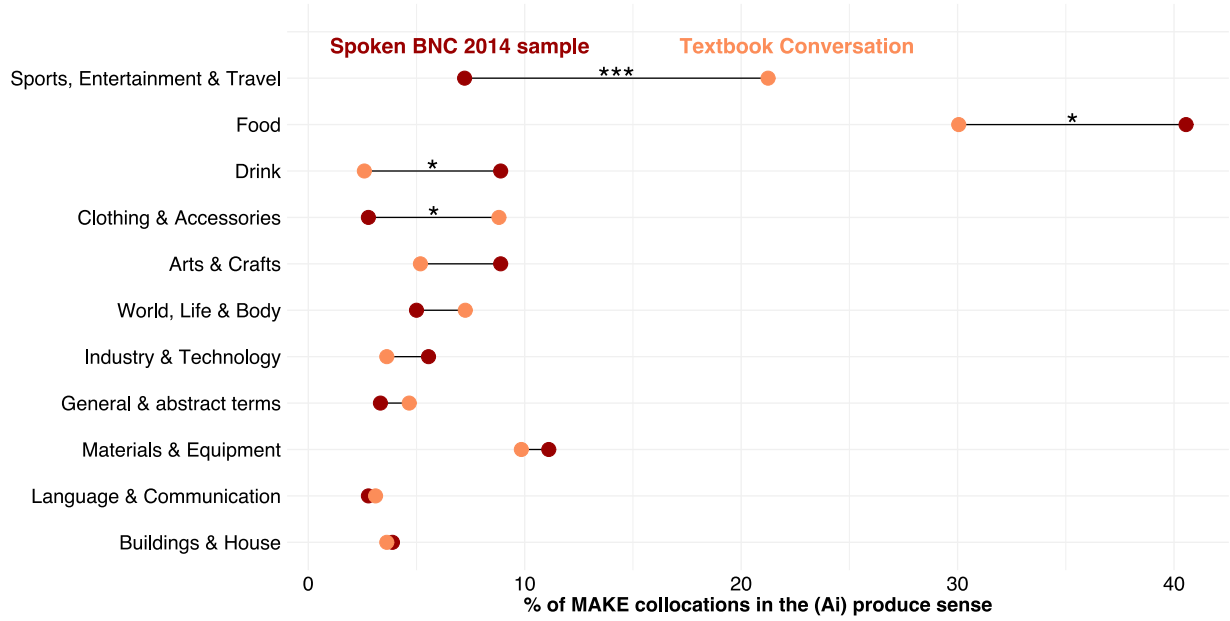


Fig. 21: Distribution of the semantic fields attributed to the noun collocates of MAKE in the (Ai) 'produce' sense in Textbook Conversation and the Spoken BNC2014 sample

In order to examine the actual lexical co-occurrences at work behind these numbers, Fig. 22 displays the most frequent object collocates. As 'produce' MAKE collocations are more frequent in Textbook Conversation than in the Spoken BNC2014 sample, relative frequencies as a percentage of all the 'produce' MAKE occurrences within each dataset are plotted and since the frequencies of the collocates follow Zipfian distributions, they have been log-transformed for better readability. Thus, the collocates on or close to the diagonal dotted line are more or less equally represented in the two sets of 'produce' MAKEs. Collocates in the upper segment are more strongly

represented in the Spoken BNC2014 sample than in Textbook Conversation, and the opposite is true of collocates displayed in the lower segment of the plot. The font colours of the collocate lemmas correspond to their attributed semantic field categories (see 5.2).

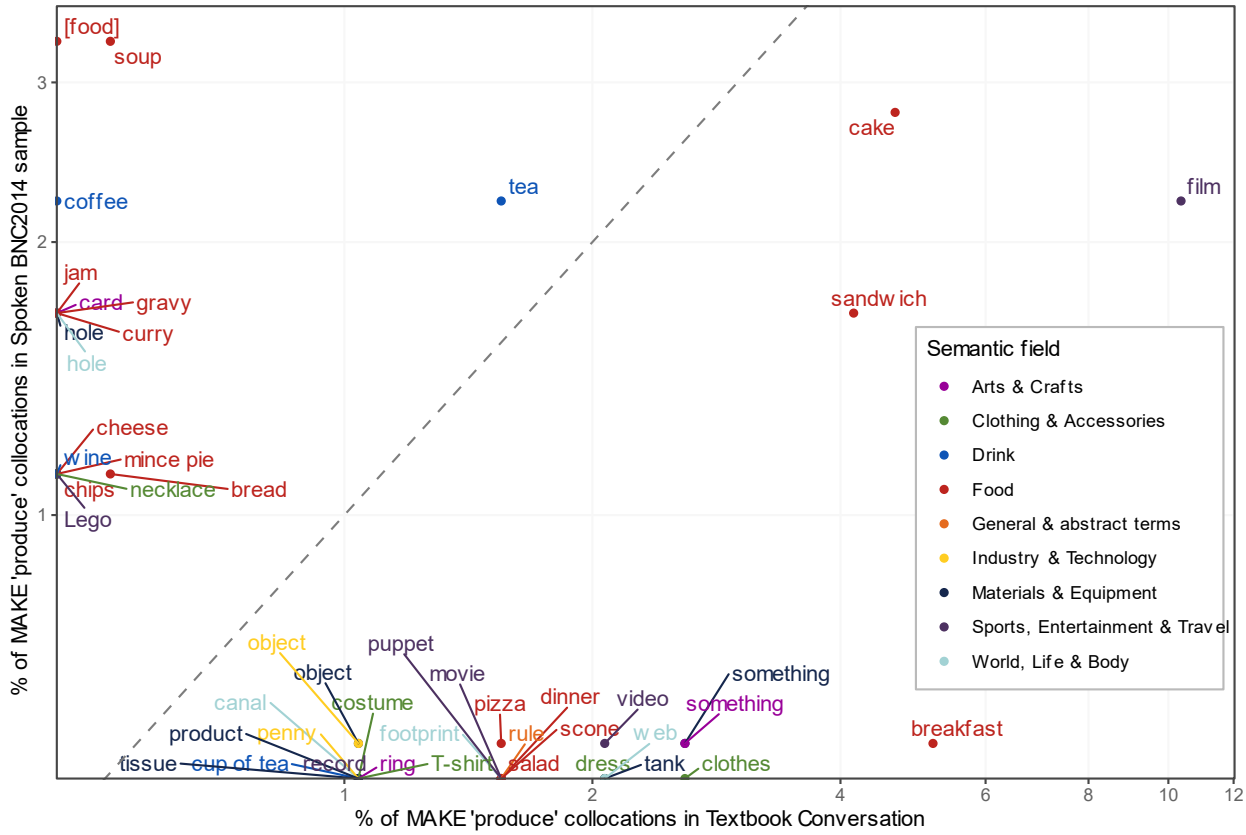


Fig. 22: Frequent collocate lemmas of MAKE in the (Ai) 'produce' sense in Textbook Conversation and Spoken BNC2014 sample

Fig. 21 shows that MAKE collocations pertaining to sports, entertainment and travel are far more frequent in textbook dialogues than in the Spoken BNC2014 sample ($p < .001$, OR = 0.29, 95% CI 0.14–0.58), whilst Fig. 22 shows that this semantic category is largely dominated by making *films*. The making of *records, movies* and *videos* is also more frequently the subject of discussions in textbook dialogues than in the reference data. Indeed, these activities are frequently discussed by the teenage characters featured in textbook dialogues, e.g., (147)–(148). Moreover, several textbook series encourage learners to produce short videos as part of their English lessons and some of the dialogues featuring these collocations aim to model the discussions that students could have when working on these video-making activities, e.g., (148).

(147) Sam: Well, do you have any other ideas? Justin: Sorry, but **I want to make a video** about Plymouth for my dad. <TEC: Access G 1>

(148) Hey! **I'm making a little film** to introduce our school to new students. Can I ask you some questions? <TEC: Green Line New 4>

By contrast, the most common semantic field associated with 'produce' MAKE collocations in the reference Spoken BNC2014 sample is 'Food', which is significantly more frequent than in Textbook Conversation ($p = .04$, OR = 1.59, 95% CI 1.01–2.49). The informal context in which the vast majority of the conversations were recorded for the Spoken BNC2014 is likely to be a major factor contributing to the high proportion of MAKE collocates relating to food and drink. Indeed, it was recommended that contributors to the Spoken BNC2014 record their conversations in quiet locations, such as at home or in cafés (Love, Hawtin & Hardie 2018: 4). Food stuffs that are frequently featured in such collocations include *chips, soups, bread, curries, mince pies, gravy, cake* and *sandwiches*, e.g., (149)–(150). The latter two are also frequently represented in Textbook Conversation.

(149) well we were gonna get a takeaway but then we just got **I made a curry and stuff**
oh nice
yeah
I love a good curry
it was good actually didn't eat for ages though it took for ages to make <BNC2014: SLDD>

The *Food* collocates found in the Spoken BNC2014 provide a fascinating window into the multicultural nature of British culinary habits; thus, the sample of 673 MAKE occurrences examined in this study includes MAKE collocations with foods as varied as *bacon baps, burgers, bhajis, chutneys, curries, fudge, mash, matzah, mince pies, toad-in-the-hole* and *Yorkshire puddings*! Confirming the British stereotype, the most frequent 'Drink' collocate is *tea*. However, it should be noted that not all MAKE + *tea* collocations refer to the brewing of Brits' favourite hot drink: in many cases, MAKE *tea* refers to preparing an (early) evening meal, e.g., (150). This common, regional use of the collocation (Smith 2018) is not featured in any of the textbook dialogues of the TEC. Conversely, MAKE + *dinner* is proportionally more frequent in Textbook Conversation than in the Spoken BNC2014, e.g., (151). Note that the use of the definite article in excerpt (151) is unusual for habitual actions: this phraseological pattern is found just once out of 38 occurrences of MAKE + *dinner* across the full Spoken BNC2014.

(150) I don't feel like cooking is my job and I have to do it
I know occasionally **make sandwiches** for me but I couldn't stand it if I was like **made sandwiches** I just think that would be horrible I just don't know it's weird or like where's my tea
yeah I've got your dinner on the table for when you come in be awful wouldn't it what I don't know I like you like **making tea** and tidying up and stuff but I think I feel guilty <BNC2014: S5YQ>

(151) I do my best to help with the housework, but it's difficult to find the time. I tidy my bedroom once a week and I sometimes take the rubbish out or help mum to **make the dinner**. <TEC: Solutions intermediate>

The most frequent food-related collocations in Textbook Conversation is MAKE *breakfast* (152). For this phraseological pattern, the absence of a definite article is accurately portrayed in the pedagogical dialogues. However, whilst also observed in the Spoken BNC2014, by far the most frequent verb collocate for *breakfast* in natural conversation is HAVE (153), rather than MAKE.

(152) Boy 1: Yeah, good idea: I can **make breakfast** for her! I never do that, ever. But now we've got a new problem: How do you **make breakfast**? <TEC: Green Line New 1>

(153) well we could do that tomorrow if you want for lunch
er I was planning on **having breakfast** for lunch tomorrow
Fair enough we could buy some brunch stuff if you want or just **have breakfast**
just **have breakfast** really I like brunch occasionally but not as a regular thing <BNC2014: SCNN>

Fig. 19 also reveals that MAKE collocations from the realm of 'Clothing & Accessories' are somewhat more frequent in textbook dialogues than in the reference conversational data ($p = .02$, OR = 0.30, 95% CI 0.08–0.86). This is likely a topic effect: items of clothing are often taught early on in EFL syllabi and this vocabulary is frequently recycled in textbook dialogues, e.g., (154)–(156). Many of these collocations are in the passive form, where either the person who designed or produced the items are the focus on the utterances (155), or the material out of which they are made (156).

(154) Your clothes look cool too. Did you use to spend a lot of money on them? I didn't use to have much money. My mother **made some of them**. And I used to share clothes with my brother. <TEC: Solutions intermediate>

(155) MEGAN: Did you watch the royal wedding? Wow, what a ceremony!
ALEX: Uhuh. I watched it for a while. Kate was beautiful, wasn't she?
MEGAN: Yes, she was. Her dress was just gorgeous. **It was made** by Sarah Burton, an English designer. Kate is such a fashion icon!
<TEC: Piece of cake 4e>

(156) Do you like my T-shirt? **It was made** from recycled plastic.
<TEC: Achievers B1>

As the interlocutors of the Spoken BNC2014 share knowledge and a spatial environment that also allows them to communicate non-verbally, the corpus features many 'produce' MAKE collocations where the collocate is not explicitly mentioned (157). Thus, the collocate lemma tag *[food]* in Fig. 21 and Fig. 22 refers to MAKE collocations that, whilst evidently referring to MAKING food of some kind, do not mention a specific collocate. In addition, vague language is much more frequent in natural conversation (158), than in the textbook dialogues which, on the contrary, are frequently very precise in their descriptions (159). As a result, textbook dialogues are characterised by high lexical diversity and many more attributive adjectives than the transcripts of authentic conversations.

(157) D'you wanna eat this?
yep yep uh thanks for **making it**
no probs
d'you wanna?
alright okay
should I get the biscuits? <BNC2014: SQ2D>

(158) or if you want them to **make something** with onion in and you just like shove them in like frozen <BNC2014: STXT>

(159) So today **we're making a lovely tomato and yoghurt sauce**. Of course tomatoes are full of vitamins, so this is a really healthy option. <TEC: Solutions Intermediate>

As already observed in Chapter 4, the differences between Textbook Fiction and the Youth Fiction reference corpus are not as pronounced as in Textbook Conversation and its corresponding reference corpus. Thus, Fig. 23 reveals just two significant differences in the distribution of the semantic fields attributed to 'produce' MAKE collocations. Note also that, although the lines plotted in Fig. 23 are longer than in Fig. 21., the range of values on the y-axis is also considerably smaller.

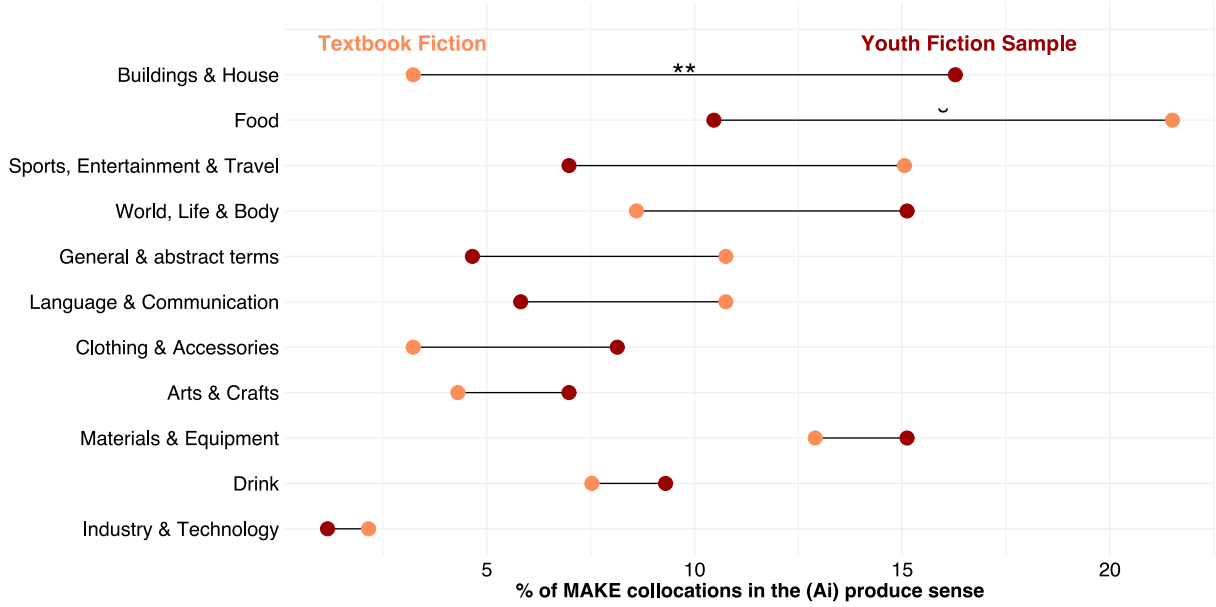


Fig. 23: Differences in semantic fields attributed to collocates of MAKE in the (Ai) 'produce' sense.

The most frequent lexical instantiations of these 'produce' collocations cannot be plotted, simply because the vast majority of collocate lemmas only occur once in each sample; thus, hapax legomena make up 95% of 'produce' MAKE collocates ($n = 77$) in the Youth Fiction sample and 75% of these collocates in Textbook Fiction ($n = 48$). Given the limited vocabulary EFL secondary school pupils are expected to have at this early stage of L2 acquisition, it is not surprising to find higher lexical diversity in novels targeted at English-speaking teenagers compared to the narrative writing of EFL textbooks.

Fig. 23 shows that the proportion of 'produce' MAKE collocates from the semantic field of 'Buildings & House' is significantly higher in Youth Fiction than in the corresponding textbook register (p = .004, OR = 5.78, 95% CI 1.53–32.57). Collocate lemmas in this semantic field include *building, carpet, chair, door, path, [zebra] pen, roof, room, settlement, shelter, steeple* and *wall.* By contrast, Fig. 23 suggests that the semantic field of *Food* may be marginally overrepresented in Textbook Fiction (p = .07, OR = .42, 95% CI 0.16–1.06). Collocates in this category include *breakfast, lunch* and *dinner*, as well as specific foodstuffs such as *pancakes*, *toast, soup* and *sandwiches.*

5.3.4 MAKE as a delexical verb

Quantitatively, Fig. 19 in 5.3.2 showed that the proportion of MAKE occurrences entering delexical collocations in Textbook Conversation is not significantly different from that observed in the Spoken BNC2014 sample (p = 0.14, OR = 0.82, 95% CI 0.63–1.07). However, a closer comparison of the specific collocates of delexical occurrences of MAKE in the dialogues and audio/video transcripts of the Textbook English Corpus with those found in the Spoken BNC2014 reveals some noteworthy patterns. Fig. 24 displays the relative frequencies of the noun object collocates as a percentage of all the delexical MAKE occurrences within each sample. Though also a frequent collocate in the Spoken BNC2014 sample, MAKE + *mistake* stands out as the most frequent delexical collocation in the pedagogical materials, whereas MAKE + *effort* is considerably less frequent in Textbook Conversation than in the reference corpus. In contrast, the most frequent delexical collocation in the reference data is MAKE + *sense*, which, although also featured, is much less frequent in the textbook data. In addition, MAKE + *difference* appears to be slightly less represented in Textbook Conversation. The fact that MAKE + *call* is considerably more frequent in Textbook Conversation is likely an artefact of the Spoken BNC2014, which does not include any telephone conversations, whereas a sizeable proportion of textbook dialogues consists of fictitious telephone conversations.

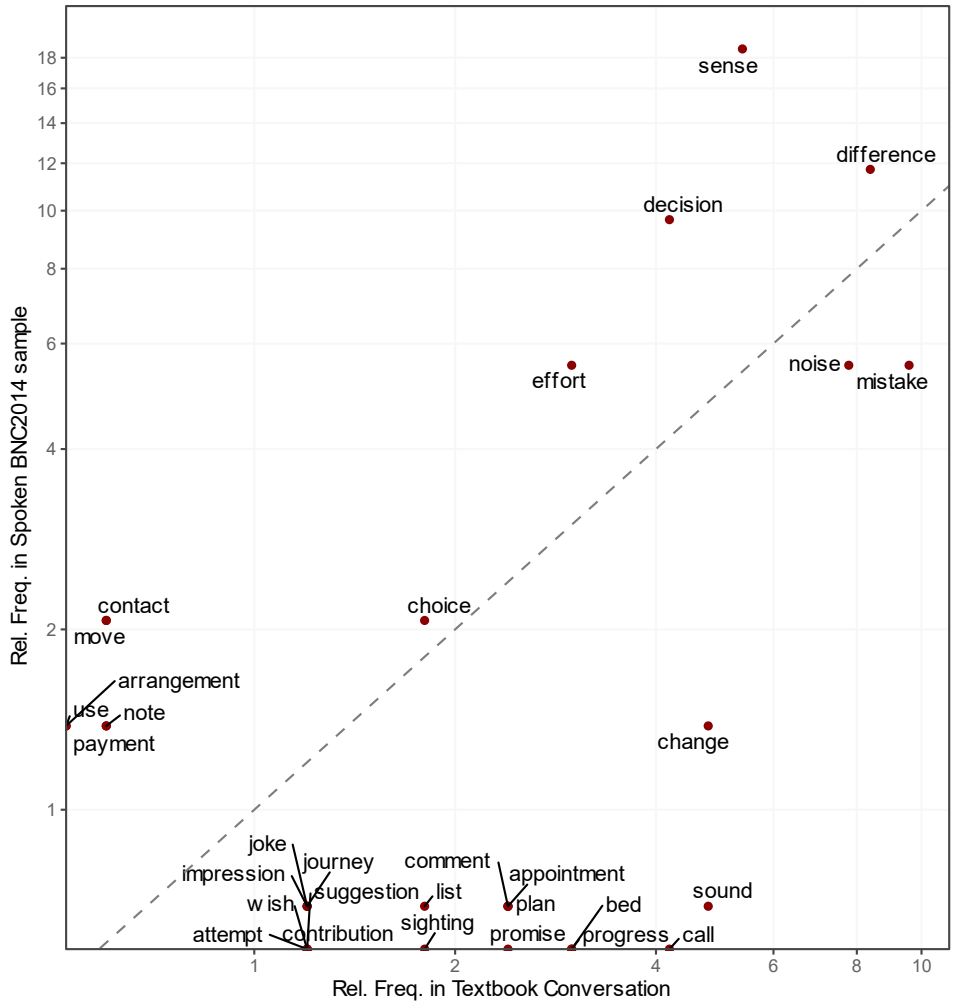


Fig. 24: Most frequent collocates of delexical MAKE in Textbook Conversation and the Spoken BNC2014

Fig. 25 suggests that delexical uses of MAKE in Textbook Fiction writing and the Youth Fiction reference corpus are more similar. Nevertheless, as in the conversation register, MAKE + *mistake* appears to be overrepresented in Textbook Fiction ($n = 13$) compared to Youth Fiction ($n = 4$). In addition, MAKE + choice is also considerably more frequent in Textbook Fiction than in the reference data, where MAKE + decision is more frequently observed. The descriptive collocations MAKE + *noise/sound/racket* are among the most common delexical MAKE collocations in both varieties of fictional writing.

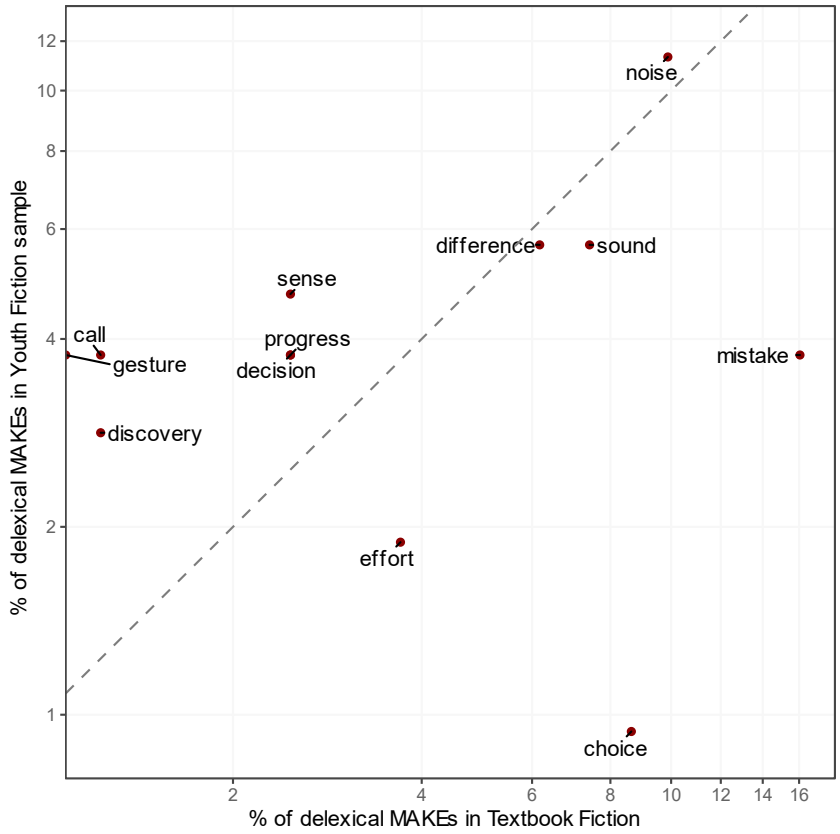


Fig. 25: Most frequent collocates of delexical MAKE in Textbook Fiction (left) and the Youth Fiction sample (right)

MAKE is known to enter into many delexical structures used to report speech (Sinclair et al. 1990: 150–151). These include MAKE + *arrangement, claim, comment, decision, promise, protest, remark, signal* and *suggestion*. In addition, MAKE is also associated with many other speech actions, e.g., *appeal, enquiry, point* and *speech* (Sinclair et al. 1990: 150–151). In their comparison of the use of MAKE in learner and native student argumentative writing, Altenberg & Granger (2001: 179–180) reported that L2 students significantly underused such 'speech' collocates of delexical MAKE; in fact, L1 students employed these collocations more than twice as frequently as L2 students.

Bearing the results from this learner corpus study in mind, it was decided to take a closer look at delexical MAKE collocates associated with speech/communicative functions in school EFL textbooks. Fig. 26 shows that the percentages of delexical MAKE collocates that refer to 'speech/communication' actions are lower in the two textbook registers than in the corresponding reference ENL corpora. However, with such low numbers involved, these differences do not reach statistical significance. Nevertheless, it is worth examining potential discrepancies at a quantitative level.

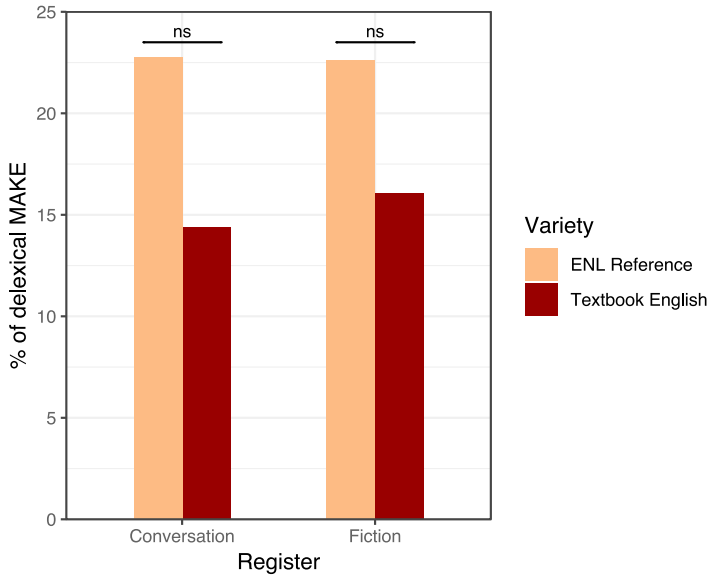


Fig. 26 'Speech/communication' delexical MAKE collocations in Textbook Conversation and Textbook Fiction (red) as compared to the reference Spoken BNC2014 and Youth Fiction samples (beige).

First, a number of 'speech/communication' delexical MAKE constructions observed in the Spoken BNC2014 are entirely absent from Textbook Conversation. These include *argument, assumption, comment, complaint, conversation, point* and *row*. Though the raw frequencies within the two samples examined are very low, it is worth noting that three of these (*argument, assumption* and *point,* e.g., (160)–(162)) were among those identified as very frequent in ENL essays and underrepresented, or not featured at all, in EFL essays (Altenberg & Granger 2001: 179–180).

(160) yes but it's difficult isn't it? it's like the whale hunting
mm
I suppose whale hunting it's easier to **make an argument** to ban it because they're endangered species
oh right yeah
whereas bulls aren't endangered <BNC2014: SPG4>

(161) and I and when you meet new new people they normally make an they **make an assumption** about you about straightaway it's human human nature you make a judgement about someone <BNC2014: S52C>

(162) I **make a valid point** you know I do just <BNC2014: SHXJ>

It might be that textbook authors perceive these delexical constructions as less worthy of teaching or more error-prone than lexical verbs such as *argue, assume* and *complain*, e.g., (163)–(165).

(163) You **argued** that we should sign a petition, but I think a flashmob would be more effective. <TEC: Access G 5>

(164) Just because I'm an American teenager, people **assume** I want to wear the same brands as JLo or Sarah Jessica Parker! That's a cliché!
<TEC: New Bridges 2[e]>

(165) I was paying by card and I didn't check the amount before I entered my PIN. Anyway, I'm sure the price ticket on the shelf was £10, but she charged me £15. I **complained** and tried to get my money back. <TEC: Solutions pre-intermediate>

Thus, we may speculate that this underuse of 'speech/communication' delexical MAKE collocates in EFL learner essays may be partially attributable to textbooks' underrepresentation of such frequent discourse structures (166)–(167). However, it should be stressed that this claim cannot be empirically backed up with the present, very sparse, data.

(166) um I mean we could book the flights now in fact maybe that's safest you reckon? Do you wanna **make er this decision** now? I just wonder cos that's quite cheap <BNC2014: SU82>

(167) do do you think it's um sort of people are **made aware** of it enough to know? You know what's going on in terms of security settings and things like that or not? <BNC2014: S2PS>

5.3.5 Phrasal verbs with MAKE

Though here, too, the absolute figures are low, it was noted in 5.3.2 that the proportional use of MAKE in phrasal verb constructions was significantly lower in both textbook registers under study than in the corresponding ENL reference corpora. Indeed, out of the 674 occurrences of the verb MAKE in the Textbook Conversation subcorpus, just eight were found to form phrasal verbs (thus less than one occurrence per textbook series!) – as opposed to 26 in the randomly collected 674 concordance lines from the Spoken BNC2014 ($p = .003$, OR = 3.33, 95% CI 1.45–8-58). Similarly, out of the 392 MAKE concordances extracted from the Textbook Fiction subcorpus, 13 were coded as phrasal verbs, whereas 29 were identified in the random sample from the Youth Fiction corpus ($p = 0.02$, OR = 2.33, 95% CI 1.15–4.96).

Phrasal verbs with MAKE are usually not introduced until the third or fourth year of English instruction. Indeed, no phrasal verb featuring MAKE was found in any beginner (Level A) textbooks. The majority are in Level D and E textbooks. In total, seven different phrasal verb types are represented in the Textbook Conversation and Fiction subcorpora (see Table 36). However, only two occur more than once, in both cases just twice, in the dialogues and narrative texts of any full textbook series (*MAKE up =comprise* in *Hi There* and *Green Line New; MAKE up =invent* in *Green Line*). One French textbook series, *Piece of Cake,* does not feature any phrasal verbs involving MAKE in its spoken language component (nor in its narrative writing but this is less surprising given that the French textbooks contain comparatively few narrative texts, see 3.3.1.4).

Table 36: Absolute frequencies of phrasal verbs with MAKE (out of 674 occurrences of MAKE each for the Conversation samples and 392 for the two Fiction samples)

	Conversation		Fiction		
Phrasal verb types	**Textbook Conversation**	**Spoken BNC2014 sample**	**Textbook Fiction**	**Youth Fiction sample**	**Examples**
MAKE it up to *sb.* (=*forgive*)	0	0	0	1	"I don't know how to **make it up to you**," he said helplessly <Youth Fiction: Ashley 2010: Chocolate Wishes>
BE made of (=*capable*)	1	0	0	2	[…] it makes you stretch your abilities and see what you're really **made of** <TEC: Access 3>
MAKE of *sth.* (=*interpret*)	0	0	0	2	Oh sir, do read the directions and see what you **make of** them <Youth Fiction: Blyton, 1943, Five Go Adventuring Again>
MAKE out (=*perceive*)	0	1	4	10	I can't quite **make it out** from my desk <TEC: Green Line New 5>
MAKE out (=*pet*)	0	1	0	0	no but it makes you go crazy like in a sexual way so you just start **making out** with whoever's around <BNC2014: SFYP>
MAKE out (=*present as*)	0	0	0	1	He had tried to make up his mind whether [...] to **make out** that he had hurt himself inside very badly <Youth Fiction: Blyton, 1943, The Mystery of the Burnt Cottage>
MAKE out to be (=*pretend*)	0	1	0	0	She's some princess [...] but I think they're **making her out to** be far worse than what she probably is <BNC2014: SAR5>
MAKE up for sth. (=*compensate*)	0	1	1	1	You're separated from your family, and so you make friends with other people to **make up for that** <TEC: Join the Team 3e>
MAKE up (of) (=*comprise*)	3	1	4	1	Its collection's **made up of** over million objects - so we won't be seeing them all today! <TEC: Achievers B1+>
MAKE up (=*face paint*)	0	1	0	2	her fear was clowns so they got John to get **made up** and and dressed as a clown <BNC2014: S57J>
MAKE *sth.* up (=*invent*)	2	17	4	6	I always thought they **made that up** to scare us <TEC: Green Line 5>
MAKE *sth.* up (=*prepare*)	1	1	0	2	I had asked him to **make up** a fire in my office <TEC: New Bridges 2e>
MAKE up (=*reconcile*)	0	2	0	1	we had a bit of a bond over that [...] yeah I think you two'll **make it up** <BNC2014: SEPP>
MAKE with (=*content*)	1	0	0	0	Oh, I'm sure we can **make with** what we've got <TEC: Green Line 2>
Total	**8**	**26**	**13**	**29**	

Both the sparsity and the high type-token ratios of MAKE phrasal verbs in Textbook Conversation (five types for eight tokens) and Textbook Fiction (five types for 13 tokens) suggest that school EFL textbook authors do not have a dedicated strategy for exposing learners to the most frequent, useful, or easily acquired phrasal verbs. Indeed, there is seemingly no agreement as to which phrasal verbs are likely to be most relevant or useful to teenage language learners. This finding echoes that of Koprowski (2005: 330) and Gouverneur (2008b: 240), both of whom concluded that

EFL textbook authors did not seem to adopt any coherent criteria for selecting the phraseological items featured in the textbooks they analysed.
Whilst this case study on the verb MAKE cannot pretend to provide a comprehensive overview of the most frequent phrasal verbs involving MAKE in ENL Conversation or Fiction, the following section aims to demonstrate that even the basic criterion of frequency can already provide textbook authors with valuable information as to which relevant phrasemes they could more prominently feature in their publications. For instance, within the 674 randomly collected concordance lines featuring the verb MAKE in the Spoken BNC2014, one phrasal verb clearly emerges as particularly prominent in everyday conversation: MAKE *sth. up* in the sense of fabricating a story (see Table 36). Indeed, it was found 17 times, whereas all other phrasal verbs within this sample of the Spoken BNC2014 were observed just once or twice (see Table 36). In addition to stories and anecdotes (168)–(169), lyrics, numbers and statistics are also frequently associated with this phrasal verb (170).

(168) he told loads of fibs in well it's not fibs at that age but yeah **he made a lot of stuff up** <BNC2014: S57G>

(169) **you've made that up** that's a load of old twaddle that never happened <BNC2014: SYX3>

(170) I think there were only about three or four actually businesses mm as far as I could tell and that they were there **making the numbers up** <BNC2014: SP2Y>

Predictably, and indeed as with all the lexical phenomena examined so far, fiction, as a register, displays a broader lexical range of phrasal verbs than conversation. In the Youth Fiction sample, *MAKE sth. up* (= *invent*) is the second most frequent MAKE phrasal verb. With nine occurrences (out of 392 MAKE occurrences), the most frequent is *MAKE out* (= *perceive*). It is used both in the visual (171) and auditory (172) senses. Strikingly, this phrasal verb is not featured in any of the narrative texts printed in the 42 textbooks under study.

(171) Then I see a messenger ride in from the east, but I can't **make out** who it is because I'm looking almost straight into the sun.
<Youth Fiction: Crossley-Holland 2001: The Seeing Stone>

(172) Daphne tried to **make out** what was said next, but people all started talking at once. <Youth Fiction: Pratchet 2001: Nation>

5.3.6 Causative MAKE[33]

In 5.3.2 we saw that causative constructions are the most frequent use of MAKE in the Spoken BNC2014 sample and the second most frequent in Textbook Conversation. As compared to the Spoken BNC2014, Textbook Conversation significantly underrepresents MAKE as a causative verb ($p = .011$, OR = 1.36, 95% CI 1.07–1.74). By contrast, no significant difference in causative usage was observed between Textbook Fiction and the corresponding sample of MAKE concordance lines from the Youth Fiction corpus ($p = .64$, OR = .92, 95% CI 0.67–1.26). Since causative constructions represent a highly frequent use of the verb MAKE, and as causative periphrastic constructions are known to cause learners difficulties in terms of both syntactic and phraseological choices (see 5.1.1), this section aims to further explore the syntactic and phraseological representations of MAKE as a causative verb in Textbook Conversation.

Table 37 displays the raw and relative frequencies of the four types of causative MAKE constructions observed in the two conversation samples. These are featured in very similar proportions and none of the differences are statistically significant. However, it is worth noting that all of the verbal MAKE causative constructions in Textbook Conversation are [X MAKE Y V_{inf}] constructions (see Table 37). This is interesting because textbook grammars place a strong emphasis on the syntactic difficulties involved in producing causative constructions in the passive voice, e.g., [X BE made $V_{to\text{-}inf}$], yet seemingly choose not to model these in textbook dialogues (for more on the representations of causative constructions in EFL textbooks and potential pedagogical implications, see Le Foll forthcoming).

[33] Causative constructions across the entire TEC corpus were further analysed in a research article to appear as: Le Foll, Elen. Textbooks and Corpus Linguistics: the case of causative constructions. In Kieran Harrington & Patricia R. Ronan (eds.), *Corpus Linguistics in the English Language Teaching Classroom - Research and Practice*. Palgrave MacMillan.

Table 37: Causative MAKE constructions in the Textbook Conversation and Spoken BNC2014 sample

Construction	Example	Freq in Textbook Conversation	Freq in Spoken BNC2014
[X MAKE Y AdjP]	The atmosphere is so oppressive and the characters are so insane, it just **made me really uncomfortable**. <TEC: New Bridges 3e>	41.28 % (71)	36.06 % (75)
[X MAKE (Y) NP]	So friendly, and they don't make you feel like you're just another tourist. They **made it a fantastic experience** for me. <TEC: Green Line New 3>	8.14 % (14)	9.62 % (20)
[X MAKE Y V_{inf}]	I didn't actually steal any money. I wanted to **make people understand** the dangers of cybercrime. <TEC: Solutions intermediate plus>	50.58 % (87)	53.85 % (112)
[X MAKE Y V_{pp}]	I can't remember exactly what it was that he said or what would be said but he would **make it known** that he was disappointed in the fact that I wasn't sticking up <BNC2014: SKPP>	0 % (0)	0.48 % (1)

In the following, we turn to the most frequent collexemes associated with these causative MAKE constructions: for [X MAKE Y V_{inf}] constructions, this means the non-finite verb, e.g., in the example from Table 37, *understand*; for [X MAKE Y AdjP] constructions, the adjective, e.g., *uncomfortable*; and for the rarer nominal constructions, the lemma of the head of the nominal phrase, e.g., *experience*. Note that verbal MAKE constructions associated with more than one verb were counted as one separate causative construction per non-finite verb slot: e.g., in (173), one instance of MAKE + STOP and one instance of MAKE + THINK were accounted for. This also corresponds to the procedure followed by Gilquin (2012: 11 fn).

(173) The sign **makes you stop and think**, doesn't it? <TEC: Green Line New 3>

Since there are considerably more causative MAKE occurrences in the Spoken BNC2014 sample than in the Textbook Conversation, Fig. 27 displays the relative frequencies of the most frequent collexemes as a percentage of all the causative MAKE occurrences within each sample. The plot can be interpreted similarly to Fig. 24. The only difference is that the colours correspond to the type of causative construction associated with each collexeme. Note that due to the Zipfian distributions of the collexemes it also features logarithmic scales.

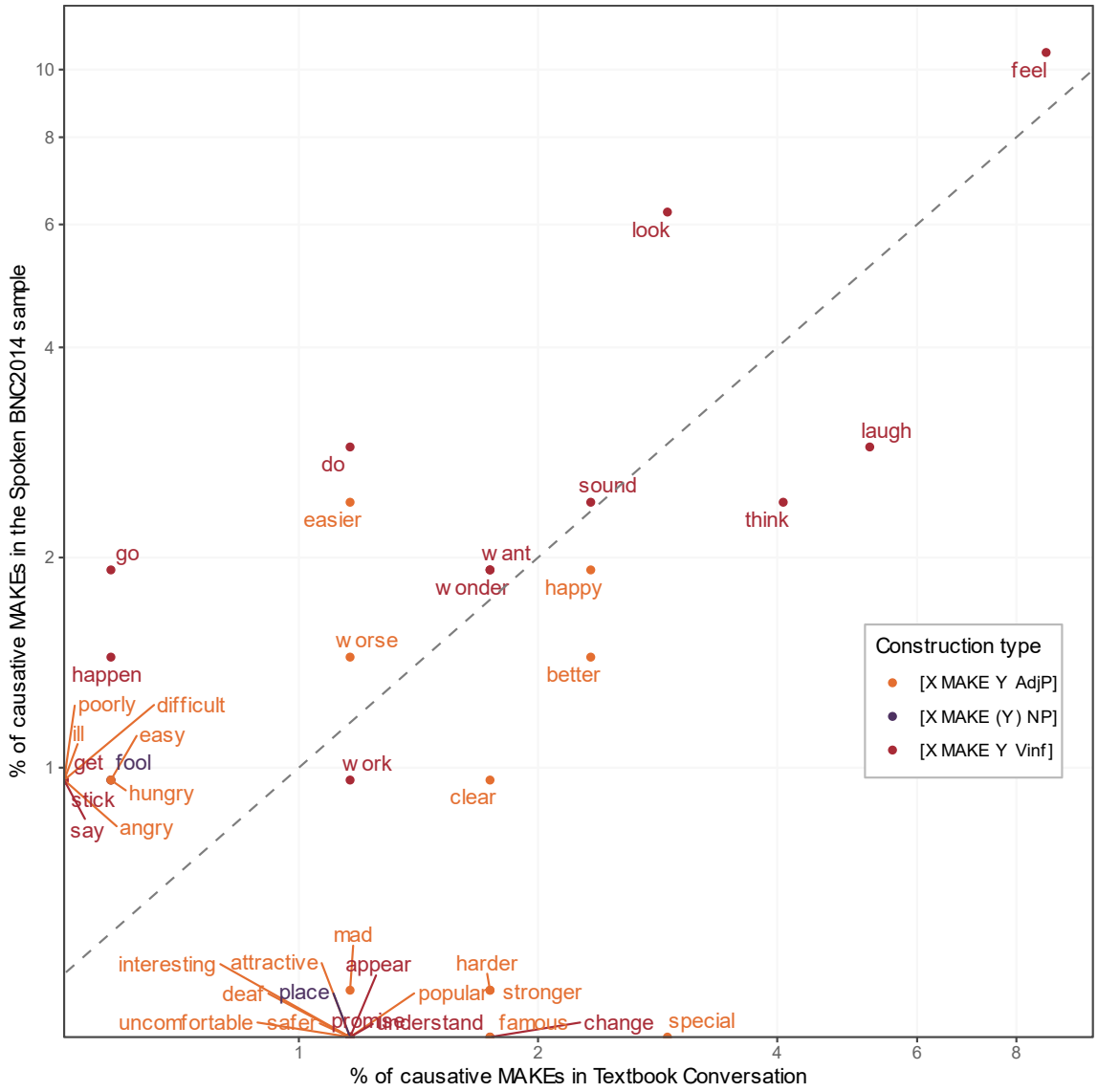


Fig. 27: The most frequent collexemes of the causative MAKE constructions of Textbook Conversation and the Spoken BNC2014 sample

Fig. 27 shows that, in both samples, *feel* is by far the collexeme most strongly associated with MAKE causative constructions. However, it is more frequently observed in the natural conversation sample ($n = 22$; 10.57% of all causative MAKE occurrences in the sample) than in the textbook dialogues ($n = 15$; 8.72%). Strikingly, both Liu & Shaw (2001) and Gilquin (2012; 2016a) report that FEEL is used significantly more frequently by EFL than ENL users in essay writing. The present results suggest that this effect may be indicative of EFL learners' limited awareness of lexical differences between different modalities and registers, rather than an overrepresentation of [X MAKE Y *feel*] in pedagogical materials.

Whilst also frequent in both datasets, Fig. 27 shows that *look* is more frequent in the Spoken BNC2014 sample than in Textbook Conversation. In the case-study chapter exploring the progressive in Textbook English, it was observed that *look* in its prototypical sense of visual perception was over-represented in Textbook Conversation (see 4.3.4.1). Here, however, [X MAKE Y *look*] constructions fulfil the

same semantic function in both datasets; it depicts appearance and, as such, *look* is most frequently followed by *like*:

(174) cos it's such a big bouffanty coat and it's not tailored it's not nicely shaped
right
it makes me look like a marshmallow <BNC2014: S575>

(175) Sarah! **You're making us look like** tourists.
So…?
Well stop it, it's embarrassing! I'm a New Yorker […]
<TEC: Green Line New 3>

As they are printed to the left of the regression line in Fig. 27, *do*, *go* and *happen* also appear to be more frequent in naturally occurring conversation than in the conversation-like texts featured in school EFL textbooks. A qualitative examination of [X MAKE Y *go*] concordances from the Spoken BNC2014 suggests that, in many cases, *go* in the non-finite verb slot of causative constructions is used as a quotative verb, e.g.:

(176) I know it's very difficult but I just don't know if like mum said more violence or retaliation against them doesn't **make them go** alright we'll stop then it makes them worse <BNC2014: SAVN>

Quotatives are very frequent in natural conversation. Indeed, Fig. 27 also shows that *say,* just like *go*, was also not observed in causative MAKE constructions in the Textbook Conversation. Qualitative analyses of the corresponding concordance lines suggest that it, too, frequently functions in this quotative sense in MAKE causatives in natural conversation, e.g.:

(177) well I think it's only fair that you pick something I don't wanna do but actually it will benefit my life massively and **make me say** oh I'm so glad you did that cos that's what's gonna happen when you watch Game of Thrones you're gonna say oh thank you for making me watch that this is amazing <BNC2014: SAG4>

By contrast, EFL textbooks traditionally neglect direct speech, usually focusing exclusively on the prescriptive rules of reported speech with person- and tense-shifting (see Barbieri & Eckhardt 2007). However, reading examples (176) and (177) also makes clear that intonation is crucial to making these quotatives intelligible. It could therefore be argued that this makes quotatives inherently unsuitable for printed dialogues; however, direct speech is not found in any of the audio and video materials of the textbooks included in the TEC either. Instead, reported speech, as it is taught in the textbooks' grammar sections is regularly modelled, even when the result is clearly not register-appropriate, e.g.:

(178) PROF: When Ruby went to school for the first time, in New Orleans, **she said that she heard** people shouting abuse. **She added that she saw** them throwing things.
TOMMY: How did the other kids react?

PROF: Not so well. She received instruction in isolation. US Marshalls had to escort her to the toilet. **She reported that she had** to eat lunch all alone in the classroom! <TEC: Piece of Cake 3e>

5.4 Conclusions

This concluding section attempts to sketch out answers to the five research questions formulated at the beginning of this chapter (see 5.1). In evaluating these results, it also reflects on the limitations of the methodology.

The first research question asked whether the verb MAKE was significantly over- or underrepresented in Textbook Conversation and Textbook Fiction as compared to ENL corpora of the same registers. It was shown in 5.3.1 that, when factoring out the overall considerably lower verb/noun ratios in Textbook English, MAKE is marginally overrepresented in Textbook Conversation as compared to the Spoken BNC2014, whilst it is seemingly underrepresented in Textbook Fiction as compared to the Youth Fiction corpus. This latter finding, in particular, was contrary to expectations. Indeed, the novels that make up the Youth Fiction corpus have a much higher lexical diversity than can be expected in the narrative texts of school EFL textbooks; yet, MAKE is the 12th most frequent verb in the Youth Fiction corpus, whereas it is ranked 16th in the Textbook Fiction subcorpus. The corresponding verb frequency lists suggest that LOOK and WANT are overrepresented in Textbook Fiction, whereas MAKE and KNOW are underrepresented.

MAKE being highly polysemous, it was deemed necessary to look at the distribution of its various meanings in order to further explore these differences at the semantic level. It was originally hypothesised that the prototypical 'produce' sense of MAKE would be overrepresented in the textbook registers, but this hypothesis was refuted (see 5.3.2). Textbook Conversation was shown to significantly underrepresent causative and phrasal uses of the verb. Idioms with MAKE were also found to be overrepresented although, as explained in 5.3.2, this is a topic-effect due to the prominence of the one idiomatic phrase: *to MAKE friends*. Textbook Fiction, on the other hand, was seen to feature considerably fewer delexical MAKE forms and, as in Textbook Conversation, phrasal verbs were also grossly underrepresented.

Sections 5.3.3 and 5.3.4 examined the noun object collocates of MAKE in the prototypical 'produce' sense and its 'delexical' usage in an attempt to answer the third research question concerning the collocates and semantic fields attributed to these collocates typically associated with MAKE in Textbook Conversation and Textbook Fiction. MAKE collocations relating to food are very prominent across all four corpora examined but may be slightly overrepresented in the reference Spoken BNC2014 corpus as a result of the convenience sampling method that perhaps favoured meal-time conversations over other conversational settings. In addition to

such topic-effects, task-effects were also identified in the textbook data, e.g., in the overrepresentation of MAKE 'produce' collocations pertaining to film and video.

The data is too sparse to provide an exhaustive quantitative evaluation of the representation of phrasal verbs with MAKE in Textbook Conversation and Textbook Fiction. However, section 5.3.5 made clear that the selection of phrasal verbs featured in school EFL textbooks is evidently not based on any of the widely accepted pedagogical criteria for selecting the lexis to be acquired, such as frequency, range, or learnability. Phrasal verbs involving MAKE seem to be strongly underrepresented in Textbook English and, even within a textbook series, no apparent effort is made to prioritise the acquisition of any specific multi-word verbs. Indeed, no lexical consistency could be detected across the 43 textbooks volumes or, even, at the publication series level. These results echo Koprowski's (2005) conclusion who found that less than 1% of lexical phrases were shared across three EGP textbooks, as well as the results of Gouverneur's (2008a: 240) study on phraseological patterns of MAKE in textbook exercises, which revealed that less than a fifth of the collocations featured were common to all textbooks of the same level (see 2.2.1.2).

Similarly, no systematic approach to introducing the most frequent and communicatively meaningful delexical MAKE collocations could be identified, even though such structures are known to be "particularly treacherous" for learners of English (Altenberg & Granger 2001: 189; see also Nesselhauf 2004). Thus, although the absolute number of delexical MAKE concordance lines examined here was relatively low, the comparative analysis (see 5.3.4) suggested that by far the most frequent delexical MAKE collocation in Spoken English, *to MAKE sense,* is vastly underrepresented in Textbook Conversation, whilst the collocation *MAKE + mistake* is considerably more frequent in Textbook English than in the reference corpora queried. Whilst it is to be expected that school-related and textbook task-relevant collocations, e.g., *MAKE + dialogue/film/list/sentences* etc., appear more frequently in the textbooks than in the ENL reference corpora, textbooks' strong focus on MAKING *mistakes* in dialogues and narrative texts, rather than, say, on *MAKING sense* or *an effort,* seems rather at odds with modern language teaching principles anchored in communicative approaches.

Moreover, and this may shed some light on the fourth and fifth research questions, it was found that delexical MAKE collocates associated with 'speech/communication' actions, e.g., *MAKE + argument/assumption/complaint/small talk*, are either entirely absent, or woefully underrepresented in the two textbook registers under study. Learner corpus studies had previously identified these collocations as underused by learners of English (Altenberg & Granger 2001: 179–180). Whilst they account for nearly a quarter of all delexical MAKE constructions in the two ENL reference corpora, their proportion was found to be significantly lower in the two textbook subcorpora.

It was argued in 5.3.4 that since many of these collocations are both frequent and fulfil particularly useful communicative functions, they ought to feature more prominently in school EFL textbooks. The same can be said of quotative MAKE causative constructions such as those found to be very rare in Textbook Conversation in 5.3.6.

In answer to the final research question some aspects of the present results suggest that textbook-based language input may indeed deprive learners of valuable exposure to frequent forms. This appears to concern especially delexical, phrasal and causative uses of MAKE which have been shown to be particularly problematic for learners of English and whose representations in Textbook English seemingly do not follow any evidence-based selection criteria. Thus, a number of constructions with particularly useful communicative functions, such as 'speech/communication' delexical MAKE collocations and quotative MAKE causative constructions, have been highlighted as conspicuously underrepresented in Textbook English (see 5.3.4 and 5.3.6).

6 A multi-dimensional description of Textbook English

Let's get started! We're on our way.
We're learning English! Let's go! Hooray!
<TEC: Green Line 1>

6.1 Introduction

In the literature review (Chapter 2) we saw that previous research on Textbook English has tended to focus on individual lexico-grammatical features. Collectively, these studies have provided us with a vast patchwork of (corpus-based) evidence demonstrating how individual linguistic features are (frequently mis-)represented in ESL/EFL textbooks as compared to various interpretations of what is often termed "real", "natural", or "authentic" English. However, the review concluded that such individual-feature studies cannot account for relevant interactions between features. In addition, we saw that potential differences between the various registers featured in English textbooks have yet to be adequately explored.

This chapter[34] and the following chapter aim to overcome these limitations by using a multi-variable statistical analysis aimed at reducing a large set of potentially relevant grammatical, lexical, and semantic features to a parsimonious set of meaningful factors of linguistic variation. Thus, the objective is to provide a more comprehensive view of the defining characteristics of Textbook English and of the linguistic variation found within school EFL textbooks. Bearing this in mind, this chapter will seek to answer the following research questions:

1. What is the extent of the linguistic variation across the major registers of Textbook English? How are the different textbook registers characterised linguistically? To what extent do the proficiency levels of textbooks interact with register-based variation? Do some textbook series show significantly more or less register-based variation?
2. To what extent do Textbook English registers differ from situationally similar, naturally occurring registers? To what extent are (some of) the observed patterns moderated by textbook series, their country of use, and/or the proficiency level of individual textbook volumes?
3. What are the defining linguistic features that characterise Textbook English registers as compared to these target language registers?

To answer these questions, Biber's (1984; 1988) multi-feature/multi-dimensional analytical framework of register variation is applied to the study of Textbook English.

[34] Selected results from this chapter have been published as: Le Foll, Elen. 2021. Register Variation in School EFL Textbooks. *Register Studies* 3(2). https://doi.org/10.1075/rs.20009.lef.

Before detailing the two possible approaches to the multi-dimensional analysis (hereafter MDA) of textbook language, 6.1.1 explains the principles behind MDA. For reasons of space, this chapter only provides a brief outline of the method. It is, however, described in detail in a number of book-length publications (e.g. Biber 1984; 1988: chaps. 5–6; Biber et al. 2004: sec. 4.4-4.5; Biber & Conrad 2019: chap. 2; Friginal & Hardy 2014). Next, 6.1.2.1 and 6.1.2.2 describe how MDA has already been successfully applied to the exploration of textbook language in English for Academic Purposes (EAP) and English L1 contexts: first, in a 'full MDA' involving biology and history university textbooks (Conrad 1996a; 2013) and, second, in an 'additive MDA' (Berber Sardinha et al. 2019) exploring US-American elementary school textbooks (Reppen 1994a; 2013). Section 6.2 then outlines how additive MDA is applied in the context of the present study.

6.1.1 Multi-feature/multi-dimensional analysis (MDA)

The MDA framework was pioneered by Douglas Biber (1984; 1988; 1995) to capture the underlying dimensions of variation across different registers of natural languages. It is based on the theoretical assumption that "differences in registers include patterns of co-occurring lexico-grammatical features" (Halliday 1988: 162), which result from texts having register-specific contexts of use and communicative goals (Biber & Conrad 2001; cf. Hymes 1984). MDA is used to reduce these large matrices of linguistic co-occurrence patterns to a few core functional dimensions of systemic, situational variation. Thus, it allows for the conceptualisation of register variation as a continuous phenomenon, which varies along multiple fundamental dimensions. It has been successfully applied to tease out register differences at different levels of granularity, e.g., between a broad range of registers as different as face-to-face conversation and official documents (e.g., Biber 1988), but also between academic writing across different disciplines (e.g., Gray 2015), or student essay writing across different levels of proficiency (e.g., Friginal & Weigle 2014).

MDA is an exploratory method and therefore makes no a priori assumptions about how the registers explored may differ from one another. As in any corpus-based analysis, in conducting an MDA, the first step consists in selecting, collecting and sampling the texts to be analysed in conjunction with the relevant metadata. The corpus ought to be representative of the variety and full range of the registers to be explored. In parallel, potentially relevant linguistic features need to be determined. At this stage, the aim is to be as inclusive as possible, so as not to omit any inconspicuously relevant features that may not have been identified in previous studies (Egbert & Staples 2019: 132). Biber's (1988) original study of spoken and written registers of English included 67 lexical, grammatical and semantic features,

ranging from first-person pronouns[35] to verbal contractions and downtoners (see Table 38 for full list). Due to the large number of features and texts involved, the features chosen are best operationalised such that they can be automatically identified and counted. This inevitably limits the types of linguistic features that can be entered in an MDA, but modern taggers are very powerful, so that the benefits of being able to count a wide range of features across very many texts largely outweigh the drawbacks (but see 6.5 for limitations).

Table 38: Linguistic features used in Biber's (1988) MDA of general English (as categorised and listed in Conrad & Biber 2013: 18-19)

Tense and aspect markers	**Prepositional phrases, adjectives, and adverbs**
1. past tense	39. total prepositional phrases
2. perfect aspect	40. attributive adjectives (e.g., *the small room)*
3. present tense	41. predicative adjectives (e.g., *the room is small*)
Place and time adverbials	42. total adverbs (except those in any other category)
4. place adverbials (e.g., *behind, downstairs, locally*)	**Lexical specificity**
5. time adverbials (e.g., *eventually, immediately*)	43. type/token ratio (in the first 400 words of each text)
Pronouns and pro-verbs	44. mean word length
6. first-person pronouns (e.g., *I, me, my, mine, we, our)*	**Lexical classes**
7. second-person pronouns (e.g., *you, your, yours*)	45. conjuncts (e.g., *alternatively, therefore)*
8. third-person pronouns (excluding *it*)	46. downtoners (e.g., *mildly, partially, somewhat)*
9. pronoun *it*	47. hedges (e.g., *almost, maybe, sort of*)
10. demonstrative pronouns *(this, those* as pronouns)	48. amplifiers (e.g., *completely, totally, utterly*)
11. indefinite pronouns (e.g., *anyone, everybody*)	49. emphatics (e.g., *a lot, for sure, really)*
12. pro-verb DO (but the algorithm listed in Biber 1988: Appendix II actually identifies DO as a main verb!)	50. discourse particles (e.g., sentence initial *anyhow, now, well)*
Questions	51. demonstratives
13. direct WH-questions	**Modals**
Nominal forms	52. possibility modals *(can, could, may, might)*
14. nominalizations (all nouns ending in *-tion, -ment, -ness, -ity)*	53. necessity modals *(must, ought, should)*
15. gerunds (participial forms functioning as nouns)	54. predictive modals *(shall, will, would)*
16. total other nouns	**Specialised verb classes**
Passives	55. public verbs (e.g., *COMPLAIN, EXPLAIN, PROMISE)*
17. agentless passives	56. private verbs (e.g., *BELIEVE, THINK, KNOW)*
18. *by*-passives	57. suasive verbs (e.g., *COMMAND, PROPOSE, RECOMMEND*)
Stative forms	58. *SEEM* and *APPEAR*
19. *BE* as main verb	
20. existential *there*	
Subordination features	

[35] This chapter uses Biber's (1988) terminology when referring to the linguistic features of the first MDAs. This means that, for example, the category of 'first-person pronouns' includes reflexive pronouns and possessive determiners (see Table 38 for details).

Table 38: Linguistic features used in Biber's (1988) MDA of general English (as categorised and listed in Conrad & Biber 2013: 18-19)	
21. *that* verb complements (e.g., *We felt that we needed a financial base.*)	
22. *that* adjective complements (e.g., *It's quite obvious that…*)	
23. WH-clauses (e.g., *I wondered what to do.*)	
24. infinitives	
25. present participial adverbial clauses (e.g., *Screaming with rage*)	
26. past participial adverbial clauses (e.g., *Given these characteristics*)	
27. past participial postnominal clauses (e.g., *the exhaust air volume required by the grid*)	
28. present participial postnominal clauses (e.g., *the currents of dissent swirling…*)	
29. *that*-relative clauses on subject position (e.g., *the papers that are on the table*)	
30. *that*-relative clauses on object position (e.g., *the papers that she thought…*)	
31. WH-relatives on subject position (e.g., *people who know him*)	
32. WH-relatives on object position (e.g., *people who he knows*)	
33. pied-piping relative clauses (e.g., *the way in which food is digested*)	
34. sentence relatives (e.g., *We waited for six hours, which was ridiculous.*)	
35. causative adverbial subordinator (*because*)	
36. concessive adverbial subordinators *(although, though)*	
37. conditional adverbial subordinators *(if, unless)*	
38. other adverbial subordinators (e.g., *insomuch as, such that, while)*	
Reduced forms and dispreferred structures	**Coordination**
59. contractions	64. phrasal coordination (NOUN *and* NOUN; ADJ *and* ADJ; V *and* V; ADV *and ADV*)
60. complementizer *that* deletion (e.g., *I think [Ø] he's gone already.*)	65. independent clause coordination (clause initial *and)*
61. stranded prepositions (e.g., *the person that I was talking to*)	**Negation**
62. split infinitives (e.g., *I want to completely convince you that…*)	66. synthetic negation (e.g., *No evidence was found…*)
63. split auxiliaries (e.g., *They have apparently sold it.*)	67. analytic negation (e.g., *That's not true.*)

Once the texts have been automatically tagged (or partially automatically tagged, cf. Le Foll 2021a: 28–29) for the chosen linguistic features, the total number of occurrences of all the features selected are counted in each text of the corpus. These raw counts are then normalised to a common denominator (e.g., 1,000 words) to enable comparisons across texts of different lengths, as illustrated in Table 39. Excerpts of the texts (179)–(181) in which an example selection of eight features were counted can be found below.

Table 39: Selected normalised feature counts (per 100 words) in three texts (see excerpts (179)-(181) below)

	Text (179)	Text (180)	Text (181)
Attributive adjectives	4.20	7.92	7.41
because	0.66	0.21	0.13
Contractions	5.97	0.00	3.84
1st person pronouns	5.34	0.00	7.41
Negation	2.12	0.43	1.85
Nominalisations	0.54	2.78	0.00
Prepositions	5.04	13.49	6.35
2nd person pronouns	4.15	0.00	3.57

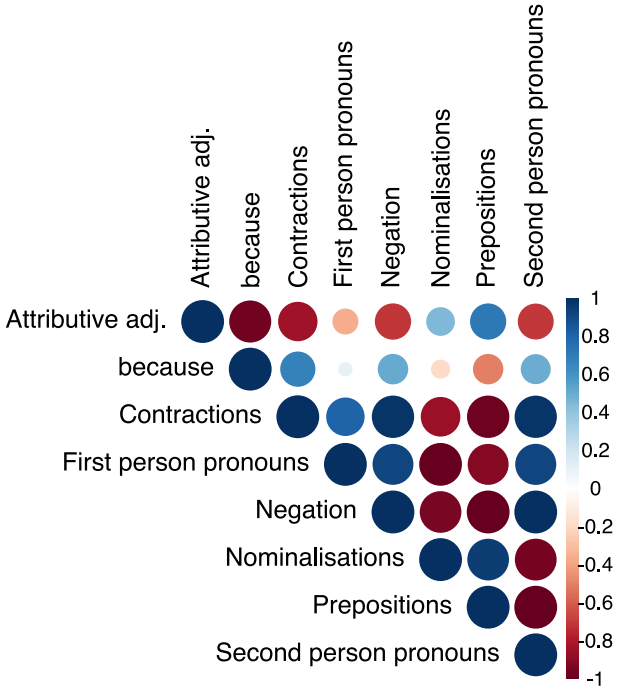


Fig. 28: Correlation matrix of the normalised counts in Table 39

A correlation matrix of all the normalised feature counts is then computed (see Fig. 28, in which the strength of the correlation between any two features is represented by the size of the circle, whilst its colour indicates the sign of the correlation). Since language features are not randomly distributed but rather according to contextual usage and communicative aims, we expect to observe many significant correlations. Indeed, Fig. 28 shows that a text with many occurrences of first and second-person pronouns is also likely to feature more negated verbs, contracted verb forms, and causative adverbial subordinators (e.g., *because*). Such positive correlations, marked in blue in Fig. 28, are frequently found in involved, spontaneous spoken conversations, as illustrated in (179), in which these features have been highlighted.

(179) **I just** did these well **I just** did these staid really laboured monologues which **you'd get** from textbooks and
yeah yeah yeah
and it was cringe cringeworthy John what **I** taught and **I'm** thinking why

didn't I think? but that**'s** that**'s** but that**'s because you're not because** that**'s** the problem **you're not** encouraged to think the the teaching language teaching industry **doesn't** encourage **you** to think it encourages **you** to use textbooks textbooks
it **no** but yeah **because** I was a newly qualified TEFL teacher **I** was obsessed with sticking to the plan
yeah
and the techniques that **I'd** been taught but they **didn't** teach **me** to use **my** knowledge and to say what do people really say in English **you** know? <BNC2014: SHJJ>

By contrast, high normalised frequencies of nominal forms ending in *-tion, -ment, -ness* and *-ity* tend to correlate negatively with the features highlighted in (179), but correlate positively with prepositions, attributive adjectives and high type/token ratios (see Fig. 28). Such clusters of features are typical of edited, information-dense texts, as illustrated in (180).

(180) **Ionesco, Eugène** özhĕn´ yōnĕs´kō, 1912-94, **French** playwright, b. Romania. Settling **in** France **in** 1938, he contributed **to** Cahiers du Sud and began writing **avant-garde** plays. His works stress the **absurdity** both **of bourgeois** values and **of** the way **of** life that they dictate. They express the **futility of human** endeavor **in** a universe ruled **by** chance. His play La Cantatrice chauve (1950; tr. The Bald Soprano, 1965) was suggested **by** the **idiotic** phrases **in** an **English language** textbook; it has become an enormously **popular** classic **of** the theater **of** the absurd. <Info Teens: factmonster.com>

Naturally, the normalised counts of the linguistic features mentioned above can also be calculated for an excerpt from a textbook dialogue (181) to compare these frequencies to those counted in "real-life English" excerpts (179) and (180). Thus, in this toy example, we can see that, although excerpt (181) purports to be spoken interaction, it features almost as many attributive adjectives as an informative text from the Info Teens (see reported frequencies for Text (180) and Text (181) in Table 39). The textbook dialogue (Text (181)) also features fewer causative subordinators, verbal contractions, negated verbs, and second-person pronouns than the conversation transcript from the Spoken BNC2014 (Text (179)). MDA facilitates these kinds of comparisons across large numbers of texts and variables.

(181) Jennifer: Hi Grandpa!
Grandpa: **Good** morning, honey!
Jennifer: What are you doing?
Grandpa: I**'m** looking at my old fairy tale book ...
Jennifer: It**'s** beautiful!
Grandpa: What**'s your favourite** tale?
Jennifer: I think the **funniest** tale is The three **little** pigs.
Grandpa: I agree with **you**! The **Big Bad** Wolf is so ridiculous!
Jennifer: Yes, it is. I like **Sleeping** Beauty too. It**'s** the most **romantic** story and Prince Charming is so handsome! <TEC: Piece of Cake 6e>

More precisely, MDA is applied to tease out the quantitative relationships – in statistical parlance referred to as the 'shared variance' – between linguistic features (variables) across a large corpus of texts. This is achieved on the basis of a correlation matrix of normalised variable counts, similar to that presented in Fig. 28, albeit much larger. The statistical method used to this effect is called exploratory factor analysis. It extracts factors that correspond to clusters of frequently co-occurring linguistic features. By definition, a factor analysis can continue to extract factors until all of the shared variance has been accounted for; indeed, once the first factor has been determined, the second factor accounts for the maximum amount of shared variance remaining, as does the third, etc. However, beyond the first few factors, additional factors are unlikely to account for more than nontrivial amounts of shared variance and may therefore be disregarded. It is up to the researcher to determine how many factors account for a sufficient amount of shared variance and can meaningfully be interpreted. In his original MDA of general English, Biber (1988) extracted seven factors, which, together, account for 51.9% of the total shared variance.

Several linguistic features contribute to, or load on, each of the extracted factors. The strengths of their relationship to a factor are captured by the factor loadings. Factor loadings thus reflect the amount of variance a feature has in common with the total pool of shared variance accounted for by any one factor. Features with a factor loading above a certain cut-off point are considered relevant contributors to the factor. Biber (1988: 87) included all features with an absolute factor loading of > 0.35 in his final model. This resulted in a final factor solution involving 60 (out of the original 67) linguistic features loading onto seven factors. This solution is summarised in Table 40, which lists the salient co-occurring features that constitute the seven factors along with their factor loadings. Note that the positive and negative signs of the loadings on any one factor serve to identify features that occur in a complementary pattern. Thus, as observed in excerpts (179) and (180), when a factor's features with positive loadings frequently co-occur within a text, those with negative loadings are, on average, also markedly less frequent (or even entirely absent), and vice versa. Features listed in brackets on Table 40 were not included in Biber's (1988) final model because they have a higher loading on a different factor and, in order "to assure the experimental independence of the factor scores" (Biber 1988: 93), each feature was only included in the computation of a single factor score.

Table 40: Features with a minimum factor loading of ±0.35 that make up Biber's (1988) seven-factor solution

Factor 1	Loading
Private verbs	.96
that-deletion	.91
Contractions	.90
Present tense verbs	.86
Second-person pronouns	.86
DO as pro-verb	.82
Analytic negation	.78
Demonstrative pronouns	.76
General emphatics	.74
First-person pronouns	.74
pronoun *it*	.71
BE as main verb	.71
Causative subordination	.66
Discourse particles	.66
Indefinite pronouns	.62
General hedges	.58
Amplifiers	.56
Sentence relatives	.55
WH-questions	.52
Possibility modals	.50
Non-phrasal coordination	.48
WH-clauses	.47
Final prepositions	.43
(Adverbs)	(.42)
Nouns	-.80
Word length	-.58
Prepositions	-.54
Type/ token ratio	-.54
Attributive adjectives	-.47
(Place adverbials)	(-.42)
(Agentless passives)	(-.39)
(Past participle WHIZ deletions)	(-.38)

Factor 2	Loading
Past tense verbs	.90
Third-person pronouns	.73
Perfect aspect verbs	.48
Public verbs	.43
Synthetic negation	.40
Present participial clauses	.39
(Present tense verbs)	(-.47)
(Attributive adjectives)	(-.41)

Factor 3	Loading
WH-rel. clauses on object positions	.63
Pied piping constructions	.61
WH-rel. clauses on subject positions	.45
Phrasal coordination	.36
Nominalizations	.36
Time adverbials	-.60
Place adverbials	-.49
General adverbs	-.46

Factor 4	Loading
Infinitives	.76
Prediction modals	.54
Suasive verbs	.49
Conditional subordination	.47
Necessity modals	.46
Split auxiliaries	.44
(Possibility modals)	(.37)

Factor 5	Loading
Conjuncts	.48
Agentless passives	.43
Past participial clauses	.42
by-passives	.41
Past participial WHIZ deletions	.40
Other adverbial subordinators	.39

Factor 6	Loading
that-clauses as verb complements	.56
Demonstratives	.55
that-relative clause on object positions	.46
that-clauses as adjective complements	.36

Factor 7	Loading
SEEM/APPEAR	0.35

The next step in an MDA involves the functional interpretation of each factor, with its co-occurrence patterns of features and their loadings, as an underlying dimension of variation. To this end, a functional micro-analysis of the individual features is conducted, seeking the shared function(s) of the clusters of features loading on each factor. Functionally interpreted factors are then referred to as 'dimensions'.

Table 43 summarises the functional interpretation of Biber's (1988) six dimensions of general English and their associated linguistic features.[36] It was derived from the linguistic features that load on each factor listed in Table 40. For instance, features with positive factor loadings on the first factor include those identified as particularly frequent in excerpt (179): first and second-person pronouns, negated verbs, and contractions. In terms of a functional interpretation, it can be said that these features are "associated with an involved, non-informational focus, related to a primarily interactive or affective purpose and on-line production circumstances" (Conrad & Biber 2013: 24). Thus, texts with a high proportion of nouns, prepositions, and attributive adjectives, as well as long words and a high type/token ratio are typical of highly informational texts with precise lexical choices, as illustrated in (180). Consequently, Biber (1988) interpreted the first factor as the 'Involved vs. Informational Discourse Dimension', whereby positive Dimension 1 scores correspond to involved texts and negative Dimension 1 scores to informational discourse.

Finally, for each text in the corpus, dimension scores for each of the dimensions identified may be computed. Before doing so, however, the normalised counts are standardised to a mean of zero and a standard deviation of one (resulting in *z*-scores) to prevent particularly frequent features from having a disproportionate impact on the computed dimension scores. This can be illustrated with a simplified example, in which a dimension has six features, with present tense, discourse particles, negation and *because* loading positively, whereas nouns and type/token ratios (TTR) load negatively. As shown in Table 41, if we simply added the normalised frequencies of the positively loading features and subtracted the negative ones to calculate the dimension scores of these three texts, we would conclude that Texts (179) and (181) are very similar to each other on this dimension. However, a closer look at the normalised frequencies presented in Table 41 reveals that, across all six features, Texts (179) and (180) are, in fact, much more similar to each other than Texts (179) and (181) are. However, because nouns are overall much more frequent than the other five features, any small relative differences in the noun counts will unduly influence the dimension scores if normalised, rather than standardised, frequencies are used.

[36] Although Biber (1988) originally extracted a seven-factor solution, he did not attempt to interpret the seventh factor because it only has one feature that loads above the pre-determined threshold.

Table 41: The computation of dimension scores on the basis of normalised frequencies

	Present tense	Discourse particles	Negation	*because*	Nouns	TTR	Dimension score
Text (179)	4.93	0.68	1.53	0.17	37.91	0.53	**-31.13**
Text (180)	4.80	0.64	1.60	0.16	41.00	0.53	**-34.33**
Text (181)	5.67	0.81	3.24	0.81	41.31	0.34	**-31.12**
Mean	*5.13*	*0.71*	*2.12*	*0.38*	*40.07*	*0.47*	
SD	*0.47*	*0.09*	*0.97*	*0.37*	*1.88*	*0.11*	

In Table 42, by contrast, the dimension scores are based on standardised frequencies (z-scores). These have been calculated on the basis of the mean and standard deviation of the normalised frequencies of each feature (see Table 41). The dimension scores thus computed in Table 42 make evident that Text 1 is, indeed, more like Text 2 than Text 3. Hence, with standardised frequencies, the features with the highest relative frequencies, here nouns, no longer exert undue influence on the dimension scores. In other words, standardised frequencies "give each feature a weight in terms of the range of its variation, rather than in terms of its absolute frequency" (Biber 1988: 95).

Table 42: The computation of dimension scores on the basis of standardised frequencies (z-scores)

	Present tense	Discourse particles	Negation	*because*	Nouns	TTR	Dimension score
Text (179)	-0.43	-0.34	-0.61	-0.56	-1.15	0.58	**-1.37**
Text (180)	-0.71	-0.79	-0.54	-0.59	0.49	0.58	**-3.70**
Text (181)	1.14	1.13	1.15	1.15	0.66	-1.15	**5.07**
Mean	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	
SD	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	

Table 43: Summary of Biber's six dimensions of English (1988)

Dimension	Description	Features
1. *Involved vs. Informational Discourse*	Low scores indicate informationally dense discourse, e.g., official documents and academic writing, whereas high scores indicate that the text is affective and interactional, e.g., face-to-face and telephone conversations.	Involved production features: private verbs, *that*-deletions, contractions, present tenses, second-person pronouns, DO as pro-verb, analytic negations, demonstrative pronouns, emphatics, first-person pronouns, *it*, BE as main verb, causative subordinations, discourse particles, indefinite pronouns, hedges, amplifiers, sentence relatives, WH-questions, possibility modals, non-phrasal coordination, WH-clauses, stranded prepositions. Informational production features: nouns, longer words, prepositions, higher type/token ratio, attributive adjectives.
2. *Narrative vs. Non-Narrative Concerns*	Works of fiction score high on this dimension, whereas official documents, academic prose and broadcasts score lowest.	Narrative concerns features: past tense, third-person pronouns, perfect aspect, public verbs, synthetic negations, present participial clauses.
3. *Explicit vs. Situation-Dependent Reference*	Low scores indicate dependence on the context, as is the case in sport broadcasts and conversations, whereas high scores indicate independence from context, e.g., academic prose and official documents.	Explicit Reference features: WH-relative clauses on object position, pied-piping relatives, WH-relative clauses on subject position, phrasal coordination, nominalisations.
4. *Overt Expression of Persuasion*	Texts with high scores explicitly mark the author's point of view and attempt to persuade, e.g., professional letters and editorials, as opposed to factual broadcasts and press reviews, which score low.	Overt expression of persuasion features: infinitives, prediction modals, suasive verbs, conditional subordinations, necessity modals, split auxiliaries.
5. *Abstract vs. Non-Abstract Information*	The higher the score on this Dimension the higher the degree of technical and abstract information, as for example in scientific discourse.	Abstract information features: conjuncts, agentless passives, past participial clauses, *by*-passives, past participial WHIZ deletion relatives, other adverbial subordinators.
6. *On-Line Informational Elaboration*	High scores on this Dimension indicate that the information expressed is produced under certain time constraints, as for example in speeches.	On-line informational elaboration features: *that* clauses as verb complements, demonstratives, *that* relative clauses on object position, *that* clauses as adjective complements.

Once dimension scores have been computed for each text, these can be compared to explore register-based linguistic variation across a corpus. Fig. 29 plots the mean Dimension 1 scores of the registers included in Biber's (1988) analysis. We see that on Biber's first 'Involved vs. Informational Discourse dimension', highly involved, spontaneously produced texts, such as telephone and face-to-face conversations, score very high and information-dense official documents and academic writing obtain low negative scores, whilst fiction scores around zero.

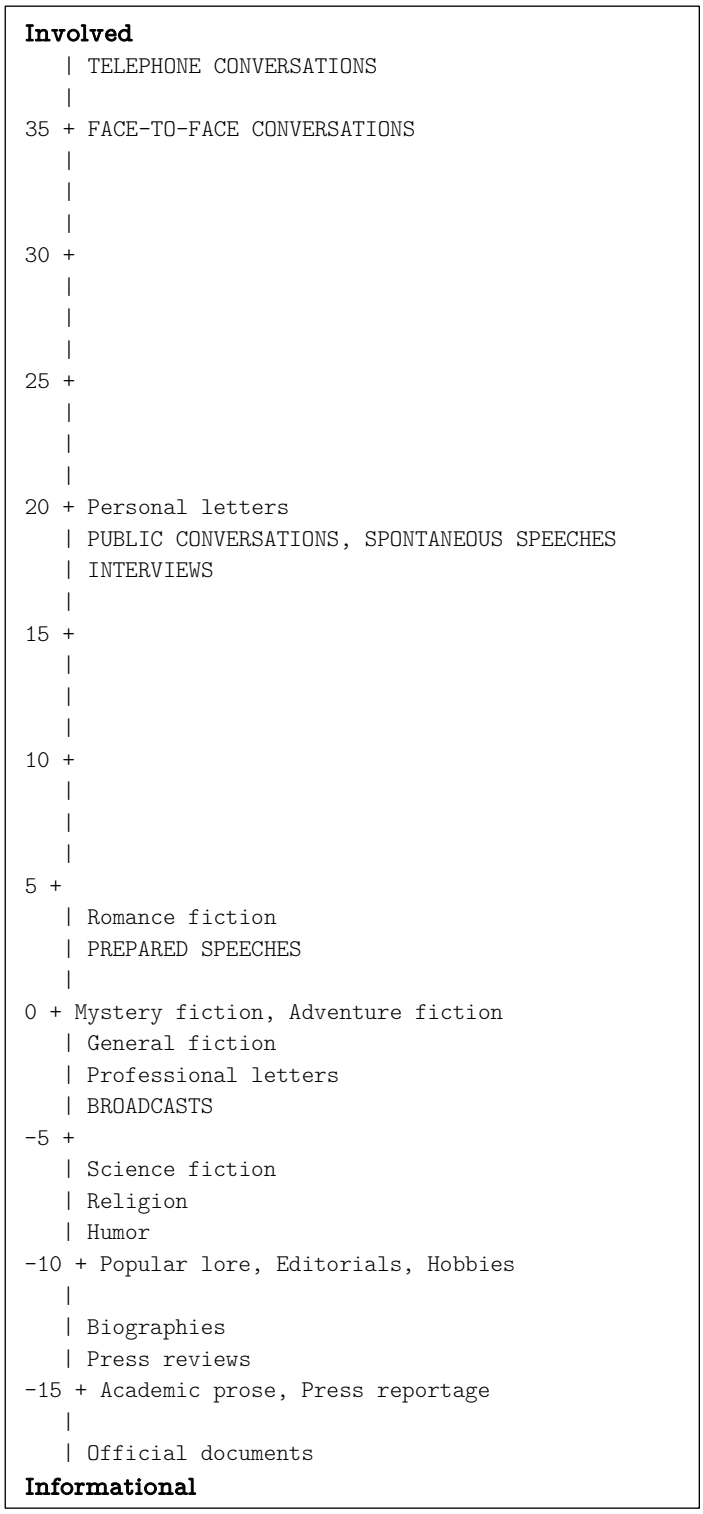


Fig. 29: Mean scores of general spoken and written registers of English on Biber's (1988) Dimension 1 (as summarised in Biber & Conrad 2019: 292)

Biber's first and the many subsequent MDAs have shown that text registers cluster in different configurations along different dimensions, thereby revealing the truly multi-dimensional nature of registers, which are characterised by several groups of linguistic features (Thompson et al. 2017: 155).

6.1.2 MDA and textbook language

Post-1988, two approaches to register variation studies applying MDA have emerged. One approach involves comparing one or more new or more specialised registers relative to the dimensions of an earlier analysis of registers (most commonly Biber's original 1988 analysis): this is referred to as 'additive MDA' (Berber Sardinha et al. 2019). The second approach consists in conducting a new, 'full MDA' following the steps outlined in 6.1.1. for an entire (new) set of registers (cf. Friginal & Hardy 2014; Egbert & Staples 2019). Given enough data, researchers can choose between these two approaches to analyse textbook language using MDA. In the following, Conrad's (1996a; 2013) investigation of biology and history university textbooks will illustrate the use of additive MDA to explore textbook language, whilst Reppen's (1994b; 2013) study of Elementary School English will serve to point to the potential of a full MDA.

6.1.2.1 Exploring Textbook English using Additive MDA

Biber's (1988) original MDA study led to the elaboration of a model of language variation in spoken and written English that can now be used for predictive purposes. With a detailed, empirical validation of its generalisation to new texts using the Brown corpus, Nini (2014; 2019) demonstrated its robustness (though see Lee 2000 for issues in replicating the six dimensions on new data). Thus, in theory at least, this means that:

> it is possible to determine how a text, corpus, or even register behaves linguistically in comparison to other registers of English. In essence, the [Biber's 1988] model represents a base-rate knowledge of English that allows the description or evaluation of other texts or registers (Nini 2019: 70).

Compared to conducting a full MDA, additive MDA approaches have the advantage of requiring considerably smaller datasets. Indeed, when conducting a full MDA, large and internally well stratified corpora are essential to be able to extract meaningful register dimensions. Where obtaining such data is not feasible, Nini (2019: 70) claims that "plotting the input corpus onto Biber's model of English can be a reasonable approximation to running a new [MDA]".

Despite this potential, relatively few studies have applied Biber's or other subsequent MDA-derived models to describe or evaluate new registers and/or varieties of English (Berber Sardinha et al. 2019). Thus far, two registers have been the focus of most additive MDAs: television registers (Quaglio 2009; Al-Surmi 2012; Forchini 2012;

Berber Sardinha 2014; Berber Sardinha and Veirano Pinto 2017) and academic registers (Atkinson 1996, Conrad 1996, 2001, 2014; Biber et al. 2002). All of these studies relied on Biber's (1988) model as their baseline.

Conrad (1996a; 2013) applied Biber's (1988) model of variation in general English to research articles and university-level textbooks in the fields of ecology and American history. She uses Biber's (1988) dimensions to compare linguistic variation between a) the two disciplines, b) professional academic writing and the pedagogical writing of textbooks, and b) all the disciplinary texts of her corpus and the English registers investigated by Biber in his original study. In the following, the focus will be on the second comparison, which concerns the specificities of Textbook English as a language variety.

Conrad's corpus comprised a total of eighty 800-word samples from 20 ecology and 20 American history research articles and ten 500-word samples from each textbook in a corpus of nine textbooks per discipline (Conrad 1996a: chap. 4 and Appendix A). She calculated the dimension scores on Biber's (1988) first five dimensions for each sample and reported mean scores for each text register (research article or textbook) and discipline (ecology or history). On Biber's (1988) first dimension, the text scores cluster at the negative end of the scale around -20 (Conrad 2001: 97). When comparing them to Biber's (1988) registers of general English, these disciplinary texts clearly form a distinct cluster. However, the additive MDA also highlights notable differences between the two academic registers. In both disciplines, the research articles feature more nouns, prepositions, and attributive adjectives and tend to have longer words, thus conveying information that is more densely packed than the textbooks. Understandably, textbooks have a more novice audience than research articles and thus include more explanations and examples, leading to less dense informational content.

In a preliminary study with a smaller corpus focusing only on biology texts, Conrad (1996b) triangulated her results with interviews with a professor of ecology who taught with the textbooks included in the corpus. Strikingly, although the professor reported that he preferred to use textbooks rather than research articles in his courses "because they were less dense and more engaging for the students" (Conrad 1996b: 314), the students were expected to develop the academic writing skills necessary to write research papers. Conrad (1996b: 302) points to a number of studies which concluded that biology students often struggle to manipulate language as their biology professors expect them to. Thus, such comparative studies of an 'input register' – here, the textbooks – and a 'target register' – in this case, the research articles – can contribute to a better understanding of the linguistic difficulties students may face and lead to more effective materials development. Conrad (1996b: 320) claims that if "the texts that [students] have been exposed to are primarily textbooks, writing an

experimental report is likely to require new ways of using language", for which students may not have been exposed to adequate models. Pursuing a similar, pedagogical aim, Zuppardo (2013) performed an additive MDA comparing the language of aircraft manuals to Biber's (1988) model and concluded that additive MDAs can support teachers and textbook authors in developing tailored materials for ESP or EAP classes by providing salient linguistic information about specialised registers.

6.1.2.2 Exploring Textbook English by conducting a full MDA

Within the MDA framework, the alternative to situating textbook language relative to other written or spoken registers of English is to conduct a new, full MDA with a corpus of textbooks or a corpus that includes textbook materials. This second methodological approach was adopted by Reppen (1994b; 2013) in her extensive study of (English L1) elementary student speech and writing. She compiled a corpus of spoken and written texts either produced or consumed by fifth-graders in the USA that included, among other texts, 10,000 words from elementary school science and social studies textbooks and 5,000 words from basal readers commonly used at that level in Arizona. Following the procedure described in 6.1.1, Reppen extracted five factors and functionally interpreted the linguistic features associated with each factor and their factor loadings to arrive at a lexico-grammatical description of the variation of registers in Elementary Student English.

Reppen was then able to both compare the registers of Elementary Student English to one another and, in a second step, to compare the dimensions of 'Elementary Student English' to those of 'Adult English' (Biber 1988). Her results show that textbooks share linguistic features typical of edited informational and non-personal uninvolved discourse (Reppen 1994a; 2013). Biber's (1988) adult model and Reppen's (1994, 2001) elementary student model share many characteristics. In both models, the first and strongest dimension depicts an oral-written continuum, reflecting production circumstances and the density of informational content. The two models also feature a second dimension pertaining to narrative vs. non-narrative discourse. There are, however, some notable differences. For instance, Reppen (2013: 196) observes that many of these reflect developmental processes because, although fifth-graders' communicative goals largely match adults', ten- to eleven-year-olds rely on a more limited set of linguistic resources to pursue the same objectives. Students presumably acquire the necessary linguistic resources to construct subtle arguments and persuade at a later stage because features associated with these communicative aims are largely absent from Reppen's dimensions of elementary school English. Moreover, Reppen (2013: 197–198) notes that some of the tasks elementary students are asked to complete at school, e.g., describing hypothetical scenarios, involve clusters of lexico-grammatical features that are not found to co-occur in adult

registers, thus – at least from a linguistic point of view – the pedagogical relevance of such school tasks may be called into question.

6.2 Method

Although, to the present author's best knowledge, MDA has yet to be applied to EFL textbook registers, both Conrad's (1996a; 2013) additive MDA involving university textbooks and Reppen's (1994b; 2013) full MDA of elementary school language have showcased the considerable potential of MDA in the context of textbook language studies. In the following, this chapter presents an additive MDA of six major Textbook English registers identified in 3.3.1.4, as compared to Biber's 1988 dimensions of spoken and written English (see 6.1.1). A full MDA of Textbook English and relevant target language registers is conducted in Chapter 7. The results of the two MDAs corroborate one another and are discussed in Chapter 8, together with the results of the case studies presented in Chapters 4 and 5.

It is hypothesised that an MDA of Textbook English will reveal which textbook registers are closest to the register they intend to represent, as well as which specific linguistic features contribute to the most striking differences between textbook registers and their intended target register. Note that this chapter and the next seek to answer the research questions outlined in 6.1 in a descriptive manner. Chapter 8 will take a more evaluative stance by discussing both the potential pedagogical reasonings behind and implications of the idiosyncrasies of Textbook English registers identified in Chapters 4 to 7.

6.2.1 Choosing a baseline MDA study

Evidently, to perform an additive MDA, the first step involves selecting an existing study whose MDA dimensions will serve as the baseline (Berber Sardinha et al. 2019: 169–170). For the present study, Biber's (1988) original MDA of spoken and written English was deemed most appropriate for five reasons.

First, it covers a broad range of registers, whereas follow-up MDAs have tended to focus on more specialised or restricted ranges of registers (e.g., University Language [Biber 2006]; academic English [Gray 2015]; blogs [Grieve et al. 2010]; online registers [Biber & Egbert 2018], etc.). In addition, Biber's 1988 model is the only MDA model whose reliability has been formally evaluated (see, e.g., Biber 1990; Biber 1992; Biber 1993b; though its validity has been more difficult to ascertain, see Lee 2000; Le Foll 2021a). Third, although Biber's dimensions have *de facto* served as a baseline for variation in all varieties of English, they were derived from British English texts (from the LOB and London-Lund corpora, except for the professional and personal letter registers, see Biber 1988: 66) and, as explained in 3.3.2.1, British English was chosen as the comparison baseline for the language of European school EFL textbooks. The

fourth reason is that relying on Biber's 1988 study allows for comparisons with other relevant MDA studies that have also applied these same dimensions (e.g., Al-Surmi 2012; Biber et al. 2002; Berber Sardinha et al. 2019; Berber Sardinha & Veirano Pinto 2017; Quaglio 2009). Finally, from a practical point of view, the availability of the Multidimensional Analysis Tagger (MAT, Nini 2014), an open-source computer programme that automatically tags and counts all the lexico-grammatical features used in Biber's (1988) analysis, eases the procedure considerably.

Biber (1988) originally extracted seven factors, though the sixth and seventh dimensions have so few features (see Table 40) that they are very difficult to interpret. Hence, in practice, only the first five or six dimensions are usually referred to in studies applying Biber's (1988) dimensions to additional registers (Conrad & Biber 2013: 39). The first six dimensions from Biber's 1988 MDA are of interest to the present investigation (see Table 43 for a summary of the dimensions and their features). The full list of the linguistic features and their loadings on each dimension can be found in Biber (1988: 102-103; Appendix II). These are the features for which each text in the corpora under study was tagged.

6.2.2 Defining text units in the TEC

The first requirement for any MDA study is to compile a text corpus representing the text categories under investigation. The design of the Textbook English Corpus (TEC) and of the three target language reference corpora, the Spoken BNC2014, Youth Fiction and Info Teens, was already described in 3.3.2.

In text-linguistic research designs, as traditionally adopted in MDA studies, the units of analysis are the individual texts within a corpus, with each text representing one observation (Biber et al. 2016: 357). Since the TEC consists of one large file per textbook volume, with each file manually annotated for register and text using a simple XML structure (see 3.3.1.3), it was possible to use a simple script to extract individual texts and remove the annotation tags from these textbook files. This process, however, resulted in very many extremely short text files (in the case of instructional texts, often just a single sentence), for which meaningful normalised feature counts cannot be computed. Linguists attempting to apply MDA to social media texts face a similar problem. To solve this issue in their multi-dimensional analysis of Twitter data, Clarke & Grieve (2017: 2) opted for binary feature frequencies (i.e., whether a feature is present or absent within a tweet) rather than relative frequencies. If, as Clarke & Grieve did, one considers a single tweet (as opposed to a thread of tweets) as a single text, this approach is very sensible because single tweets have, by corpus linguistic standards, a very small maximum character limit (currently 280 characters) and as a result, relative frequencies would largely depend on tweet length. The case of textbook texts, however, is much more complex: whilst many textbook texts are as short as a tweet (e.g., task instructions, short

rhymes), countless others run well over 1,000 words (e.g., short stories, news articles, transcript of a dialogue). Indeed, defining text units in school EFL textbooks is a particularly challenging task. Numerous possibilities arise. Up until now, entire textbook volumes have often been conceived as single texts. However, as mentioned earlier, such an approach entirely ignores the variety of text registers encountered within a single textbook volume. A second approach might consider all the texts of one register found within a chapter or unit of a textbook volume to constitute one text. In some cases, this may be justified because texts within a textbook unit will often be thematically related and may therefore form a coherent whole; however, this will depend on the textbook series and is not always consistent across an entire textbook series, either (cf. Le Foll 2020c).

In addition to the problem of defining text units, the great variety of text lengths encountered in school EFL textbooks must also be considered. Short texts may not present enough opportunities for many linguistic features to occur. In other words, even if a feature is not particularly rare, a text may simply happen to be too short to feature it. In many corpus-linguistic studies, it is often tacitly assumed that normalising counts of occurrences somehow solves this problem (see 7.2.4). In the case of zero counts, however, it evidently does not. This is easily illustrated by imagining a short informative textbook text totalling 100 words that might feature 20 nouns and six present tense verbs but not a single adverb or relative clause. If we normalise these counts to 1,000 words, we are implying that a longer version of this informative text would feature 200 nouns, 60 present tense verbs and still zero adverbs and zero relative clauses! As this example makes clear, the minimum text length must therefore be determined on the basis of the frequency of the least frequent linguistic feature to be counted in an MDA. In order to carry out an additive MDA based on Biber's (1988) model, however, the type/token ratio variable must be calculated on the basis of the first 400 words of any text[37] (Biber 1988: 238-239); thus it made sense to take this number as the minimum text length for the present additive MDA.

In light of both the great variety of text lengths encountered in school EFL textbooks and the fact that the majority are under 400 words, shorter texts within each textbook volume and register were collated into longer text files. This means that, for example, a number of short, consecutive instructional texts from any one textbook volume were combined until a total word count of at least 400 words was reached. This is standard practice in many MDA studies. On this Lee (2000: 226) writes: "[…] in cases where texts are relatively homogeneous within the genre, joining together several texts to

[37] It has long been established that type/token ratios must be calculated on the basis of text samples of equal text length as this lexical diversity measure is highly sensitive to text length (e.g., Brezina 2018: 58). Biber (1988) chose to calculate this type/token ratio of the first 400 words of each text of his corpus and, if using Biber (1988) as the base rate model in an additive MDA, this procedure should be adhered to in order for the results to be comparable.

form one large 'text' may be justified." For the texts of the TEC, concatenation was performed sequentially within each textbook volume so that short files from within a chapter/unit or across directly adjacent chapters/units were grouped together. Hence, the collated text files also correspond to the progression that the learners are expected to make. This process resulted in the exclusion of Poetry & rhyme texts from thirteen volumes, Fiction texts from seven volumes, and Informative texts from two volumes because the texts of these registers within these volumes did not total to at least 400 words. Following these data preparation steps, 1,949 textbook text files were created. Table 44 shows the dispersion of these texts across the six selected textbook registers. They are collectively referred to as the texts of the TEC in the rest of this chapter.

Table 44: Distribution of Textbook English Corpus (TEC) texts processed in this chapter

Textbook registers	Number of texts	Number of words[38]
Conversation	529	512,587
Fiction	285	241,512
Informative texts	364	304,695
Instructional texts	647	585,049
Personal correspondence	88	69,570
Poetry & rhyme	37	26,445
Total	**1,949**	**1,739,858**

6.2.3 Tagging and counting features

A marginally modified cross-platform version of the MAT (Nini 2014)[39], an open-source programme that aims to replicate the Biber Tagger, was used to automatically tag and count all the linguistic features included in Biber's (1988) final six dimensions of general spoken and written English. The validity and reliability of the MAT as compared to the Biber Tagger has been demonstrated in Nini (2014; 2019). The programme builds on the grammatical analysis and part-of-speech tagging of the Stanford Tagger (Toutanova & Manning 2000; Toutanova et al. 2003) with a series of regular expressions designed to match the feature descriptions from Biber's (1988: Appendix II) original publication. The MAT not only computes the normalised counts for all of Biber's features (and several additional features, though they are not used in this additive MDA), it also automatically standardises them based on the mean

[38] As counted by a slightly modified version of the MAT (Nini 2014).

[39] Many thanks to Stephanie Evert for improving the efficiency of the MAT and to Peter Uhrig for subsequently modifying the script for it to call on an updated version of the Stanford tagger (v. 3.9.2 with bidirectional model).

and standard deviations from Biber (1988: 77), and ultimately computes all six dimension scores for each text.

The MAT outputs three tab-separated files: 1) normed counts per 100 words for each feature per text, 2) *z*-scores based on Biber (1988) for each feature in each text, and 3) the six computed dimension scores for each text. Thus, for the purposes of a full MDA using the same linguistic features as Biber, the first table may be used, whereas the third table is of most interest to carry out an additive MDA. However, the first and second table contain valuable additional information to interpret the results of an additive MDA.

6.2.4 Computing the mean dimension scores for the new registers

We have seen in 6.1.1 that, to compute dimension scores, normalised counts must be standardised to avoid frequent features from having a disproportionate influence on the model (see Tables 41 and 42). Standardisation involves scaling the normalised counts as standard deviation units. Thus, *z*-scores for each variable in each text are calculated as follows:

(182) `(Normalised frequency in text - mean normalised frequency of corpus) / standard deviation of corpus`

Consequently, texts whose normalised count for any one variable is equal to the variable mean have a *z*-score of 0. Positive *z*-scores indicate that a feature occurs more than on average across the corpus, and negative *z*-scores indicate below average normalised counts. In an additive MDA, however, the mean *z*-scores and standard deviations of each feature are not taken from the corpus under study, but rather from the reference corpus from which the original model was derived. Thus, the second table that MAT outputs is based on the means and standard deviation values reported in Biber (1988: 77).

6.2.5 Computing dimension scores for additional reference corpora

In theory, conducting an additive MDA makes it possible to compare "new" registers to Biber's (1988) "old" general English registers without resorting to any additional reference corpora. However, in order to better answer RQ2 and RQ3, three target language reference corpora are also mapped onto Biber's (1988) dimensions for comparison with the registers of the TEC. Both theoretical and methodological reasons justify this additional step.

Starting with the methodological rationale, whilst Nini (2014; 2019) demonstrated the overall reliability of the MAT Tagger, his analyses pointed to small differences in the feature counts as compared to the original 1988 version of the Biber Tagger. In addition, the present study makes use of a slightly modified version of the MAT that relies on a more up-to-date version of the Stanford Tagger (Toutanova & Manning

2000; Toutanova et al. 2003) as the first layer of part-of-speech tagging (see 6.2.3), for which no such reliability tests were conducted. It goes without saying that the results of comparisons are more likely to be robust if exactly the same tools and methods are used to identify and count the features of both the Textbook English Corpus and any other corpora with which comparisons are to be made.

From a pedagogical and linguistic perspective, although the registers included in Biber's (1988) model can provide useful comparison points for school EFL textbook registers, any differences observed, say between Biber's (1988) fiction registers and EFL textbook fiction, could potentially be due to different target readerships. Indeed, the fiction subcorpora of the Lancaster-Oslo-Bergen Corpus of British English (LOB) predominantly contain samples from literature aimed at an adult readership, rather than children or teenagers. Additionally, the corpora from which Biber's (1988) model was derived consist of transcripts of spoken material recorded between 1953 and 1987 (London-Lund; Svartvik & Quirk 1980) and texts published in 1961 (LOB; Johansson, Leech & Goodluck 1978). Modern EFL textbooks, however, can reasonably be expected to reflect more recent language change, especially in informal, spoken language. Consequently, in the second half of the present results section (6.3.2), the dimension scores of the texts extracted from the three target language reference corpora (the Spoken BNC2014, the Youth Fiction and the Info Teens, see 3.3.2) are processed with the same modified version of the MAT for comparison with the Conversation, Fiction, and Informative texts of the TEC respectively. This 'comparative additive MDA' focuses on Biber's (1988) first three dimensions, which are identified in the analysis of intra-textbook linguistic variation (6.3.1) as the most relevant to the study of the language of school EFL textbooks.

6.2.6 Comparing dimension scores

To compare different registers on any one of Biber's (1988) dimensions, the mean dimension scores of all the texts of a register are usually compared to each other. Such comparisons have typically been tested and quantified using ANOVAs and coefficients of determination (e.g., Biber 1988: 95; Biber et al. 2004: 64; Gray 2015: 216; Berber Sardinha & Veirano Pinto 2019: 6), or with nonparametric Kruskal Wallis ANOVAs (e.g., Shakir 2020). More recently, the use of predictive Discriminant Function Analysis (DFA) as a post-hoc analysis method has also been proposed to verify the robustness of dimensions as predictors of register (e.g., Crossley, Allen & McNamara 2014; Crossley, Kyle & Römer 2019; Veirano Pinto 2019). However, a crucial assumption of both ANOVAs and DFAs is that the data points be independent of each other (cf. Gries 2015b; Winter 2019: chaps. 14–15; on the consequences of using DFA on non-independent data, cf. Mundry & Sommer 2007). In the context of the present additive MDAs, and, indeed, in many, if not most, corpus linguistic studies, however, this assumption is not met. In fact, in the case of the TEC, each textbook series has largely been written by the same group of authors, following the

same publisher guidelines. They are thus not truly independent. Similarly, the Youth Fiction and the Info Teens consist of several samples from any one book or web domain (see 3.3.2.2–3.3.2.3) which means that not all of these texts can be said to be truly independent data points.

Consequently, linear mixed-effects models were computed using the R package *lme4* (Bates et al. 2015). Following Barr et al. (2013) and many others, maximal random-effect structure models that included by-series/by-source varying intercepts and slopes were originally computed; however, for the models of the intra-textbook linguistic variation (6.3.1), this resulted in singular model fits. Singularity is typically either due to random effects having (near) zero estimated variance or to some random effects being (almost) perfectly correlated with one another. Variance-covariance matrices of the random effects were therefore computed and matrices of correlations among the estimated random effects were plotted to examine these possible causes (e.g., Fig. 30 for Dimension 2 where each point on each scatterplot represents a textbook series). These showed that these maximal models result in some almost perfect correlations between specific textbook registers and the by-series random effects. For instance, as seen in Fig. 30, in the maximal model used to predict Dimension 2 scores, the by-series random effects are almost perfectly negatively correlated with those of the Instructional register (-0.98). This means that textbook series that have a higher random intercept also have a lower random slope for the Instructional register. Since the reference group for the registers is Conversation, the intercept represents the estimated Dimension 2 scores for Conversation texts of the TEC: in other words, textbook series with higher estimated Dimension 2 scores for their dialogues also tend to feature instructional texts with lower estimated Dimension 2 scores.

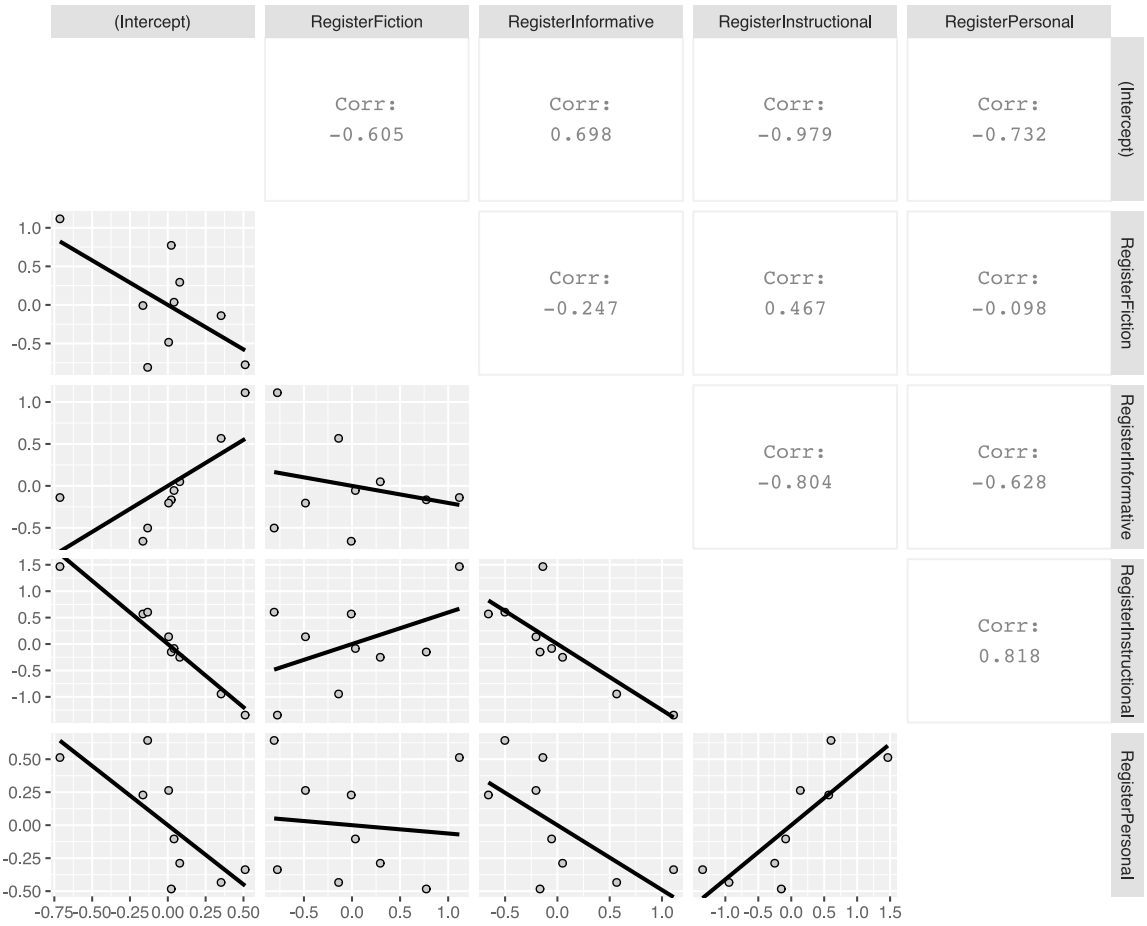


Fig. 30: Matrix of correlations among the estimated random effects of the maximal model: lmer(Dim2 ~ Register + Level + Level*Register + (Register|Series))

It thus becomes clear that some of these random slopes for the different textbook registers provide redundant information: if the random effect for one register is almost perfectly correlated with another, there is arguably no need to include them both. Whilst it is entirely feasible that some of these random effects are indeed almost perfectly correlated (near negative perfect correlations, in particular, may simply point to generally better register distinctions in some series than in others), this redundancy in the models makes it difficult to estimate the effects reliably. Several solutions have been proposed to tackle singularity in mixed-effects models (cf. Bolker 2020). Having compared these, the most satisfactory solution for the purposes of this study was to simplify the mixed-effects structures to include by-series intercepts only.

To estimate the relationships between textbook registers and the dimension scores on any one dimension, a baseline model was first fitted with a random effect structure consisting of by-series varying intercepts to account for the non-independence of texts from within one textbook series. In all models reported on in this chapter, the dimension scores are the outcome variable. Textbook register and textbook level are modelled as fixed level predictors. In addition, their two-way interaction term is also fitted, since it can be hypothesised that, as the proficiency of learners increases, the dimension scores of textbook texts within a register may move closer to their target

language equivalents. For instance, upper-intermediate fictional texts from textbooks may be more like teenage or young adult fiction than a short story printed in a beginner textbook. If this were true, we would expect dimension scores for some registers to increase as learners are expected to become more proficient, whilst they may decrease for other registers. Thus, on Biber's (1988) first dimension, advanced textbook dialogues may score higher than beginner ones, thereby more resembling naturally occurring conversation, whereas we may hypothesise that informative texts in advanced textbooks are likely to score lower on Dimension 1 than those targeted at beginner or intermediate learners. For data sparsity reasons, a subset of the data that excluded the textbook register Poetry & rhyme was entered in these mixed-effect models since several textbook volumes did not include any poems or songs longer than 400 words that could therefore be entered in the MDAs (see 6.2.2), and thus models involving this register failed to converge.

Using the *anova* function in base R, the baseline model (a) consisting only of the random by-series intercepts was compared to models with the same random effect structure and register as a fixed effect (b), textbook proficiency level as a fixed effect (c), both register and level as fixed effects (d), and with both variables as fixed effects and their two-way interaction (e). These model fits were compared by means of likelihood ratio tests by comparing the full models against the corresponding models without the effects of interest using likelihood ratio tests (cf. Baayen 2012: 675–676).

(a) lmer(DimScores ~ (1|Series))
(b) lmer(DimScores ~ Register + (1|Series))
(c) lmer(DimScores ~ Level + (1|Series))
(d) lmer(DimScores ~ Register + Level + (1|Series))
(e) lmer(DimScores ~ Register + Level + Register*Level + (1|Series))

To compare the dimension scores of Textbook Conversation, Fiction, and Informative texts with the three corresponding target language reference corpora in 6.3.2, a second type of linear mixed-effects model was computed. In these models, the random effect structure consists of varying by-Source intercepts, where 'Source' corresponds to a factor variable with nine textbook levels corresponding to each textbook series for the TEC, 300 book levels for the Youth Fiction, 14 web domain levels for the Teens Info corpus, and one level for the Spoken BNC2014. These levels have been chosen as the best-available proxies to capture the variation inherent to each (group of) author(s)/editor(s). In these models ((f)–(j)), the fixed effects are Corpus type (Textbook vs. Target Language Reference), Register (Conversation, Fiction, and Informative texts) and their two-way interactions.

(f) lmer(DimScores ~ (1|Source))
(g) lmer(DimScores ~ Register + (1|Source))
(h) lmer(DimScores ~ Corpus + (1|Source))
(i) lmer(DimScores ~ Register + Corpus + (1|Source))
(j) lmer(DimScores ~ Register + Corpus + Register*Corpus + (1|Source))

Model diagnostic plots were inspected to check the assumptions of linearity, homogeneity of variance, and the normal distribution of residuals of the model (i.e., the differences between the observed and fitted values). In addition, observed and estimated values were plotted for additional visual checks of the final model fits. Note that, for reasons of space, most of these model comparisons and visualisations are not printed but they may be reproduced using the data and code provided in the Online Appendix 6.2–6.3.

In all model summaries reported in this chapter and the next, the confidence interval (CI) ranges reported are 95% confidence intervals. The R^2_{marginal}-values summarise the predictive power of the fixed effects only, whilst $R^2_{\text{conditional}}$-values summarise those of both the fixed and random effects. The latter were computed using the R package *sjPlot* (Lüdecke 2020) on the basis of the procedure outlined in Nakagawa et al. (2017). The degrees-of-freedom (df), which represent the number of independent observations in each group and are necessary to estimate these *p*-values, were estimated using the Kenward-Roger method (as recommended in Luke 2017). The estimators of relative contrast effects between each register under study were computed using the default parameters of the *emmeans* package (Lenth 2020). Hence, here, too, degrees-of-freedom (df) were calculated using the Kenward-Roger method (Luke 2017). Further, *p*-value adjustment followed the Tukey method.[40] For these, too, the confidence level reported is 0.95.

6.3 **Results**

Section 6.3.1 attempts to answer the first research question (RQ1) formulated at the beginning of the chapter, namely: *What is the extent of the linguistic variation across the major registers of Textbook English?* In the following, this is referred to as 'intra-textbook variation'. To do so, the scores of the textbook registers on Biber's (1988) dimensions are first compared to each other. Large within-register dispersions in these dimension scores are further explored and examples of salient features that contribute to strikingly low or high scores are discussed in context. This section also seeks to find out whether some textbook series show significantly more or less register-based variation and whether the different proficiency levels of the textbooks of the TEC (significantly) interact with register-based variation. Hence, this section begins with an exploration of intra-textbook linguistic variation.

[40] Tukey's HSD (honestly significant difference) test was chosen for these post-hoc analyses because it does not make any strong distributional assumptions and operates on relatively conservative estimates for unequal sample sizes. It computes pairwise comparisons of all the means and calculates the smallest significant difference between them taking into account the cumulated Type I error level. It therefore does not result in a loss of test power in spite of the multiple comparisons it makes (Rasch et al. 2014: 29–30).

This is followed, in 6.3.2, by a more fine-grained comparison of the dimension scores of three key textbook registers (Conversation, Fiction and Informative texts) to those of three comparable target language corpora with the aim of investigating RQ2: *To what extent do textbook registers differ from situationally similar registers encountered outside the classroom?* Finally, the results of this additive MDA also provide first answers to RQ3 which seeks to pinpoint the key linguistic features which most contribute to these differences. This last research question, however, is explored more comprehensively in Chapter 7.

6.3.1 Intra-textbook linguistic variation

6.3.1.1 Intra-textbook variation on Biber's Dimension 1

Fig. 31 displays the scores of the six textbook registers on Biber's (1988) first, 'Involved vs. Informational Discourse', dimension. In Biber's original model of general spoken and written English, 84.3% of the variation in Dimension 1 scores of the individual texts from his corpus is explained by their register membership (Biber 1988: 126–127). This dimension is thus a very powerful predictor of register variation in general English. Moreover, its involved/oral/verbal versus informational/literate/nominal opposition has almost universally emerged as the strongest and most stable predictor of variation in many subsequent MDAs carried out on a range of languages and domains (Biber 2014). As was to be expected, on this 'Involved vs. Informational' dimension, Conversation scores highest ($\bar{x}$ = 15.75, SD = 7.89), followed by Personal correspondence ($\bar{x}$ = 9.62, SD = 6.81) and Fiction ($\bar{x}$ = 5.03, SD = 8.29). The lowest scores are found in the Informative ($\bar{x}$ = -5.26, SD = 7.53) and Instructional ($\bar{x}$ = -4.69, SD = 4.60) registers (see Fig. 31). However, no significant difference between these latter two textbook registers could be ascertained on this first dimension (see Table 67).

Fig. 31 shows that the TEC scores on Dimension 1 are fairly normally distributed within each textbook register, except for Poetry & rhyme, which is due to the fact that only 37 texts from this register could be entered into the MDA (see Table 44). As a result, all statistical analyses presented in the following sections exclude this register (see 6.2.6).

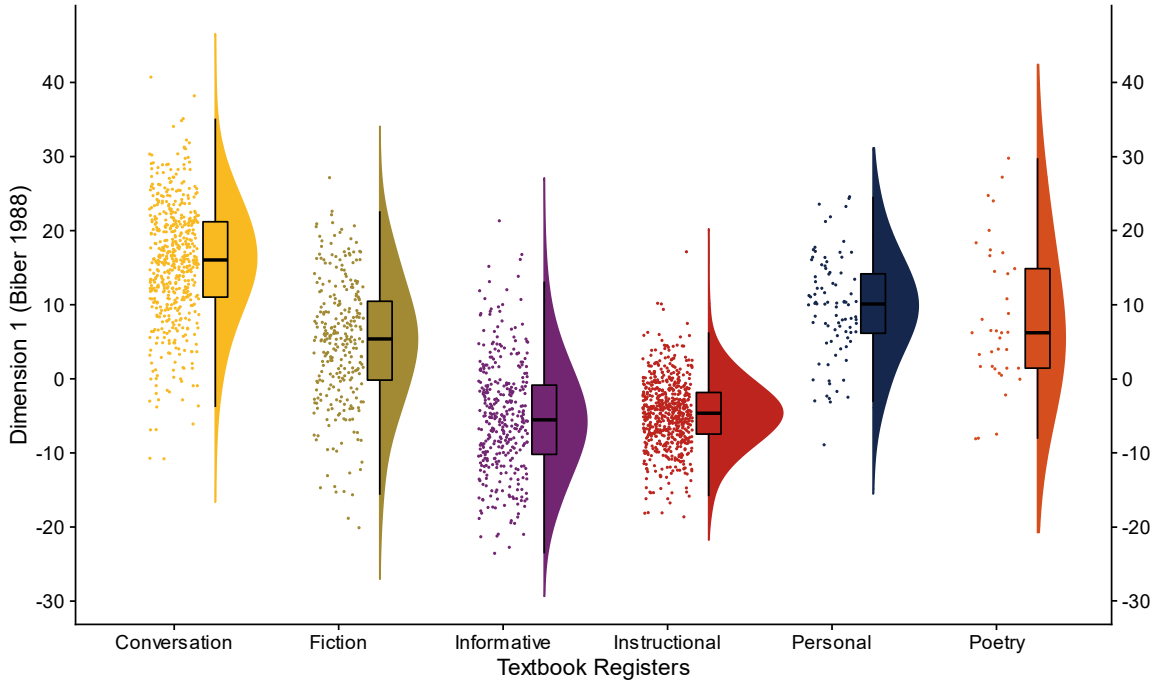


Fig. 31: Distribution of texts of the TEC on Biber's (1988) Dimension 1

As illustrated in Fig. 31, textbook register is evidently a strong predictor of Dimension 1 scores. A simple model featuring only register as a fixed effect and by-series varying intercepts already accounts for some 66% of the variance in Dimension 1 scores ($R^2_{marginal} = 0.63$, $R^2_{conditional} = 0.66$). Although model comparisons revealed that the proficiency level of textbooks is also a significant predictor of Dimension 1 scores ($\chi^2(4) = 52.27$, $p < 0.001$, as compared to the baseline model), its predicting power is very weak ($R^2_{marginal} = 0.03$, $R^2_{conditional} = 0.08$).

Table 66 summarises the final linear mixed-effects model computed for Biber's (1988) Dimension 1 scores with register, level and their two-way interaction as fixed effects, and random by-series intercepts. The reference levels are Conversation for register and A for the proficiency level of the textbooks (corresponding to the first year of secondary school, see Table 3). Thus, the coefficient estimate for the intercept (16.34) represents the estimate mean Dimension 1 score for Textbook Conversation texts in beginner (level A) textbooks. All other mean estimates on Table 66 (and all other subsequent model summary tables) represent the values that need to be added (or subtracted) from the intercept mean value to obtain the score estimates for the other textbook registers and proficiency levels. For example, the estimated Dimension 1 score for Textbook Informative texts from level A textbooks is equal to -2.61 which is the result of the sum of 16.34 (intercept) and -18.95 (Informative register). This value is very close to the actual observed mean value of -2.83 (SD = 7.52). Since all the mixed-effects models in the present chapter and the next also include Register*Level interactions, these also need to be accounted for when calculating estimated dimension scores. Thus, to find out the estimated Dimension 1 score of Level E Fiction textbook texts, it is necessary to calculate the sum of four coefficients:

16.34 (intercept), -6.96 (Fiction register), -1.27 (Level E) and -5.64 (Fiction*E interaction) which totals to 2.47 and is also close to the observed value of 3.06 (SD = 8.51).

Table 45: Summary of the model: lmer(Dim1 ~ Register + Level + Level*Register + (1|Series))

Predictors	*Estimates*	*CI*	*p-value*
(Intercept) [Conversation, Level A]	16.34	14.41 – 18.26	**<0.001**
Register [Fiction]	-6.96	-9.44 – -4.47	**<0.001**
Register [Informative]	-18.95	-21.54 – -16.36	**<0.001**
Register [Instructional]	-21.42	-23.23 – -19.62	**<0.001**
Register [Personal]	-5.29	-9.03 – -1.56	**0.006**
Level [B]	0.62	-1.12 – 2.36	0.485
Level [C]	-0.28	-2.01 – 1.45	0.753
Level [D]	-2.19	-3.94 – -0.44	**0.014**
Level [E]	-1.27	-3.25 – 0.71	0.208
*Register [Fiction] * Level [B]*	-4.09	-7.30 – -0.88	**0.013**
*Register [Informative] * Level [B]*	-1.80	-5.07 – 1.46	0.279
*Register [Instructional] * Level [B]*	0.21	-2.19 – 2.60	0.866
*Register [Personal] * Level [B]*	-2.82	-7.56 – 1.91	0.243
*Register [Fiction] * Level [C]*	-4.09	-7.31 – -0.86	**0.013**
*Register [Informative] * Level [C]*	-0.61	-3.74 – 2.52	0.703
*Register [Instructional] * Level [C]*	1.20	-1.17 – 3.57	0.321
*Register [Personal] * Level [C]*	-1.53	-6.34 – 3.29	0.534
*Register [Fiction] * Level [D]*	-6.06	-9.19 – -2.93	**<0.001**
*Register [Informative] * Level [D]*	-1.37	-4.48 – 1.73	0.386
*Register [Instructional] * Level [D]*	2.21	-0.17 – 4.60	0.069
*Register [Personal] * Level [D]*	-0.46	-5.49 – 4.56	0.856
*Register [Fiction] * Level [E]*	-5.64	-8.87 – -2.40	**0.001**
*Register [Informative] * Level [E]*	-4.45	-7.71 – -1.18	**0.008**
*Register [Instructional] * Level [E]*	-0.24	-2.83 – 2.35	0.857
*Register [Personal] * Level [E]*	-0.53	-5.58 – 4.52	0.836
	Random Effects		
σ^2	41.29		
$\tau_{00\ Series}$	4.60		
ICC	0.10		
N_{Series}	9		
Observations	1912		
Marginal R^2 / Conditional R^2	0.647 / 0.682		

Fig. 32 presents a visualisation of the model summarised in Table 66: Dimension 1 scores as predicted by the model are plotted as red triangles, whilst the actual, observed scores for each text are represented as grey dots. In addition, Fig. 32 also

serves as a reminder of categories for which there is only sparse or no data; for instance, the textbook series *Piece of Cake* (POC) and *Solutions* do not have a Level E textbook in their series, and some textbook series contain very few or no fictional texts at certain levels (see 6.2.2). These aspects need to be taken into consideration when evaluating the results of the following analyses.

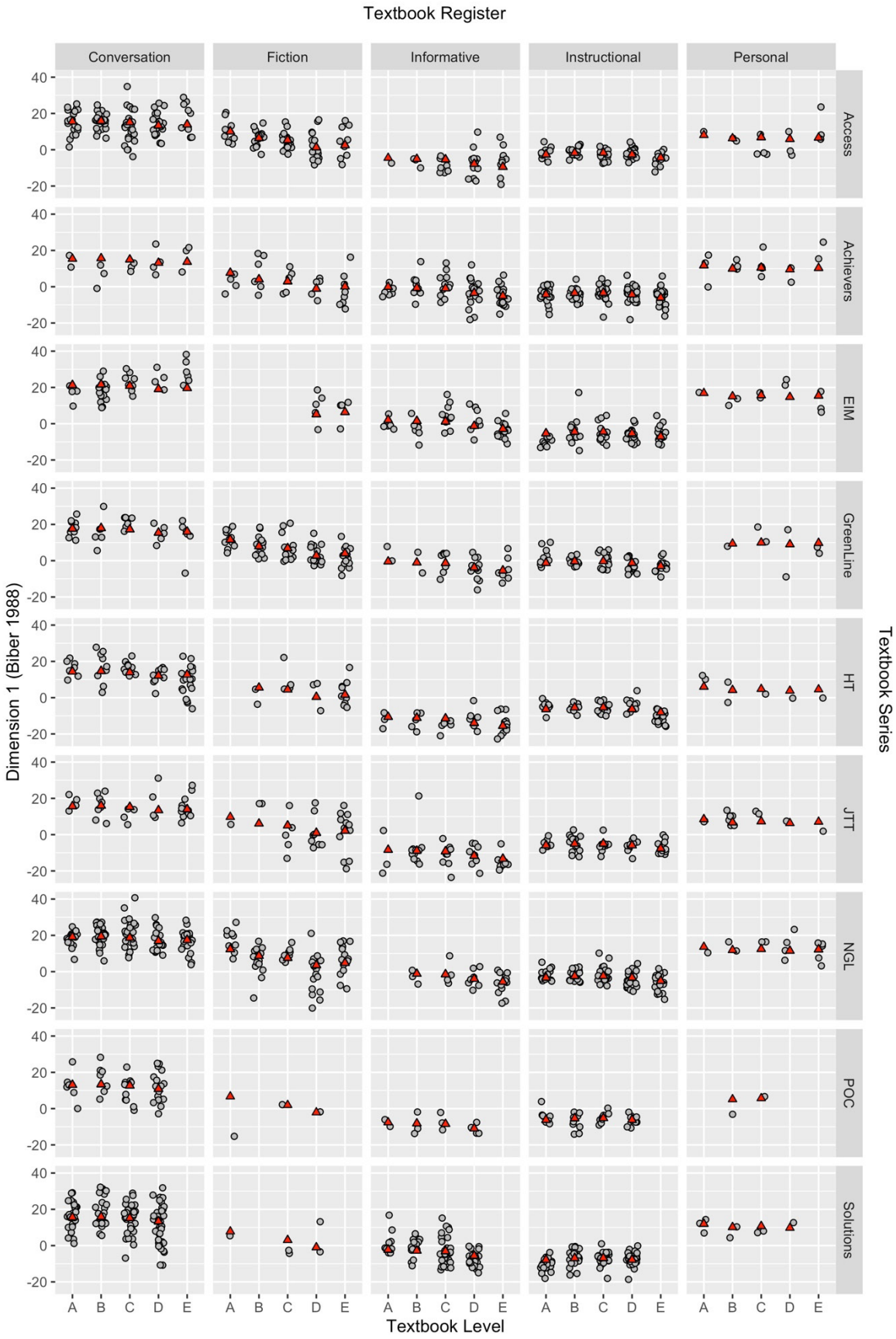


Fig. 32: Observed (grey circles) and predicted (red triangle) Dimension 1 scores across textbook register, proficiency level and textbook series. Predicted values as computed by the model summarised in Table 66. (Le Foll 2021. Zenodo. Retrieved on 7 May 2021. http://doi.org/10.5281/zenodo.4732323).

Even at this earlier stage, however, the model estimates in Table 66 confirm the need to examine school EFL textbook language under the lens of register. Indeed, register appears to have a much larger impact on the choice and frequencies of the linguistic features that constitute Dimension 1 (see Table 40) than the proficiency level of the textbook, or the linguistic idiosyncrasies of its authors (as, admittedly imperfectly, captured in the random by-textbook-series intercepts). The model estimates in Table 66 also reaffirm that textbook level only has a minimal influence on the Dimension 1 scores of textbook texts. In fact, as will be discussed further, the significant effect of textbook proficiency level on Dimension 1 scores is entirely driven by its interactions with Fiction, which tends towards very marginally lower Dimension 1 scores as the proficiency level of the textbooks increases. In addition to the small effect sizes (as shown by the estimate coefficients), this finding must be approached with caution because Fiction is known to be rather poorly captured by Biber's (1988) first dimension. Indeed, fiction consists of (frequently alternating) dialogues and narrative passages so that novels with a high proportion of dialogues will inevitably score high, whilst those with longer descriptive passages will score lower (cf. Biber & Finegan 1994; Egbert & Mahlberg 2020). Whilst it may be tempting to conclude that beginner textbooks feature more dialogue-heavy fictional writing than more advanced textbooks, it is also worth remembering that the data for the textbook Fiction register is rather sparse for several textbook series (see Fig. 32). A far more relevant factor is likely to be the fact that fictional texts printed in beginner textbooks feature many more present tense verbs simply because the past tense has yet to have been introduced.

Thus, whilst all five interactions between the Fiction register and the textbook proficiency levels are significant, the effect of textbook level on Dimension 1 scores remains very small (see Table 66). This is why the model's estimated mean values for the Dimension 1 scores of the TEC texts displayed in Table 46 have been averaged across the five textbook levels. Echoing the results displayed in Fig. 31, Table 67 confirms that the estimated register means for Dimension 1 are all significantly different from each other ($p < .001$), except for the Informative-Instructional contrast. The relatively large standard errors associated with the Personal register is due to the fact that, as also illustrated in Fig. 32, there are relatively few Personal correspondence texts in the TEC. This small sample size is also reflected in the relatively large confidence intervals of the mean Dimension 1 score for the Personal correspondence register (see Table 46).

Table 46: Estimated register means of Dimension 1 scores (averaged across all textbook levels)

Register	*mean*	*SE*	*lower CL*	*upper CL*
Conversation	15.71	0.81	12.7	13.94
Fiction	4.78	0.87	16.4	2.94
Informative	-4.89	0.85	15.0	-6.70
Instructional	-5.04	0.81	11.9	-6.79
Personal	9.35	1.0	35.3	7.24

Given these results, we can, at first blush, conclude that textbook authors do make different, register-based linguistic choices when crafting the texts featured in school EFL textbooks. Further, the distribution of scores on Biber's (1988) original Dimension 1 – with Textbook Conversation scoring highest and Textbook Informative writing at the bottom of the scale – conforms to our expectations based on previous MDA studies.

Table 47: Estimated differences between mean Dimension 1 scores for each TEC register pair (averaged across all textbook levels)

Contrast	*Estimate*	*SE*	*df*	*p.value*
Conversation - Fiction	10.93	0.505	1936	**<.0001**
Conversation - Informative	20.6	0.478	1933	**<.0001**
Conversation - Instructional	20.75	0.394	1933	**<.0001**
Conversation – Personal correspondence	6.36	0.763	1930	**<.0001**
Fiction - Informative	9.67	0.564	1936	**<.0001**
Fiction - Instructional	9.82	0.487	1936	**<.0001**
Fiction – Personal correspondence	-4.57	0.812	1930	**<.0001**
Informative - Instructional	0.15	0.452	1929	0.9974
Informative – Personal correspondence	-14.23	0.795	1928	**<.0001**
Instructional – Personal correspondence	-14.38	0.748	1928	**<.0001**

Table 48 helps shed light on the linguistic features which contribute most to the textbook register variation captured on this first dimension. It displays the mean *z*-scores for all the features that load on Biber's (1988) Dimension 1. Features that contribute to positive dimension scores are visualised in blue, whilst those that have negative loadings are listed in red. *Z*-scores above 1 or below -1 have been highlighted as particularly salient: they correspond to features whose mean values are at least one standard deviation above or below Biber's (1988) general English corpus mean. Individual *z*-scores highlighted in blue contribute to higher Dimension 1 scores, while those in red contribute to lower scores. Table 48 shows that several factors contributed to Textbook Conversation texts obtaining comparatively high scores on this dimension, including high relative frequencies of contractions, first- and second-

person pronouns, *it* pronouns, discourse particles[41], general emphatics[42], negation, and WH-questions.

Table 48: Mean *z*-scores of the features that load on Biber's (1988) Dimension 1

Dim 1 feature	Conversation	Fiction	Informative	Instructional	Personal	Poetry
Private verbs	0.19	0.23	-0.39	0.58	0.06	0.00
that-deletion	0.30	0.00	-0.30	-0.20	0.10	0.10
Contractions	1.56	0.47	-0.19	-0.46	1.13	1.23
Present tense verbs	0.26	-0.66	-0.66	-0.81	-0.15	0.10
2nd person pronouns	1.52	0.35	0.05	2.55	0.94	1.19
DO as a pro-verb	0.12	-0.29	-0.35	0.20	-.09	-0.29
Analytic negation	1.15	0.61	-0.20	-0.69	0.79	0.55
Demonstratives	0.63	-0.03	-0.33	-0.63	-0.15	-0.06
General emphatics	1.22	0.23	0.63	-0.57	1.18	0.32
1st person pronouns	1.25	0.56	-0.49	-0.91	2.02	1.84
pronoun *it*	1.02	0.52	0.10	-0.63	0.90	0.42
BE as a main verb	0.47	-0.58	-0.88	-1.73	0.06	-0.04
Causative sub.	0.16	0.12	0.29	-0.53	0.46	-0.39
Discourse particles	1.04	0.11	-0.36	0.01	-0.11	-0.39
Indefinite pronouns	-0.47	-0.29	-0.54	-0.68	-0.46	-0.23
General hedges	0.37	0.17	-0.31	-0.36	0.13	-0.05
Amplifiers	0.14	-0.13	0.08	-0.89	0.19	-0.79
Sentence relatives	0.54	0.34	2.04	0.58	0.54	0.13
WH-questions	1.66	1.52	-0.15	1.35	-.01	-.03
Possibility modals	0.92	0.48	0.22	0.39	0.30	0.75
Non-phrasal co-ord.	0.68	0.84	0.48	-0.13	0.24	0.34
WH-clauses	0.59	0.88	0.27	4.96	0.66	0.42
Final prepositions	0.15	0.18	0.45	-.07	-0.12	-0.20
Nouns	1.78	1.16	2.78	2.40	1.24	1.51
Average word length	-1.39	-1.16	0.03	-0.24	-1.34	-1.61
Prepositions	-1.76	-1.16	-0.54	-0.69	-1.42	-1.75
Type/token ratio	-1.00	-0.25	0.60	-1.31	.01	-0.88
Attributive adj.	-1.09	-0.89	0.02	-0.68	-0.80	-0.98

[41] The discourse particles included in this variable are *well, now, anyhow, anyway* and *anyways*. For full details on the operationalisation of all the linguistic variables used in this additive MDA, see Nini (2014).

[42] General emphatics are operationalised as any occurrences of *just, really, most, more, real* + ADJ, *so* + ADJ, DO + V, *for sure, a lot* and *such a* (see Nini 2014).

In addition, Textbook Conversation features low standardised frequencies of prepositions and attributive adjectives, as well as shorter than average words and low type/token ratios. Since these are features with significant negative loadings, these negative *z*-scores contributed to high scores. Table 48 further reveals that Conversation is the only textbook register with a positive mean *z*-score for demonstrative pronouns. Textbook Conversation also features comparatively more occurrences of *that* deletion, present tense verbs, DO and BE as main verbs, hedges[43] and possibility modals than the other five textbook registers.

Textbook Informative texts are characterised by comparatively few occurrences of present tense verbs, pronouns, questions, and negation, which all contribute to this register's low Dimension 1 scores ($\bar{x}$ = -5.26, SD = 7.53). Additionally, texts in this register feature the highest average word length among textbook registers and the highest type/token ratio. It is also the most nominal of textbook registers with a mean noun frequency of 27.9 per 100 words (SD = 4.80).

Although instructional texts feature very many second-person pronouns and WH-questions and have the lowest type/token ratio of all registers, all of which contribute to higher scores on Biber's Dimension 1, on average, dimension scores for the Instructional register are also characteristically low ($\bar{x}$ = -4.69). Compared to the other five textbook registers under study, they display the least intra-register variation (SD = 4.60). The low scores for Instructional texts are due to an absence of first-person pronouns, very few other pronouns aside from second-person pronouns, few contractions, negated verbs, and few occurrences of BE as a main verb. Among the features with negative loadings, the relatively high proportion of nouns also contributes to these low scores.

6.3.1.2 Intra-textbook variation on Biber's Dimension 2

As shown in Fig. 33, the distinctions between the different textbook registers on Biber's (1988) Dimension 2 are considerably less pronounced than on the first. Nevertheless, as expected on this narrative dimension, Textbook Fiction scores highest ($\bar{x}$ = 2.53, SD = 2.83), whilst Textbook Conversation scores lowest ($\bar{x}$ = -3.17, SD = 2.05).

Fig. 33 excludes a clear outlier in the Poetry register (Dimension 2 score = 16.66). This data point corresponds to the lyrics of three songs from the textbook Solutions Elementary which were collated to reach the 400-word threshold: Nina Simone's *Ain't Got No, I've Got Life,* Alicia Key's *Fallin'* and *Take the last train to Clarksville* by The Monkees. The remarkably high score on this second dimension is the result of

[43] Hedges are *maybe, at about, something like, more or less, sort of / kind of* + DT/QAN/CD/JJ/PRP$/WH (see Nini 2014).

the repeated phrase *Ain't got no* in the first song, in which all 50 instances of *got* were erroneously tagged by the MAT (see 6.2.3) as past tense verbs and of 18 occurrences of *no* which were (correctly) identified as synthetic negation. Both of these features load positively on Biber's (1988) narrative dimension.

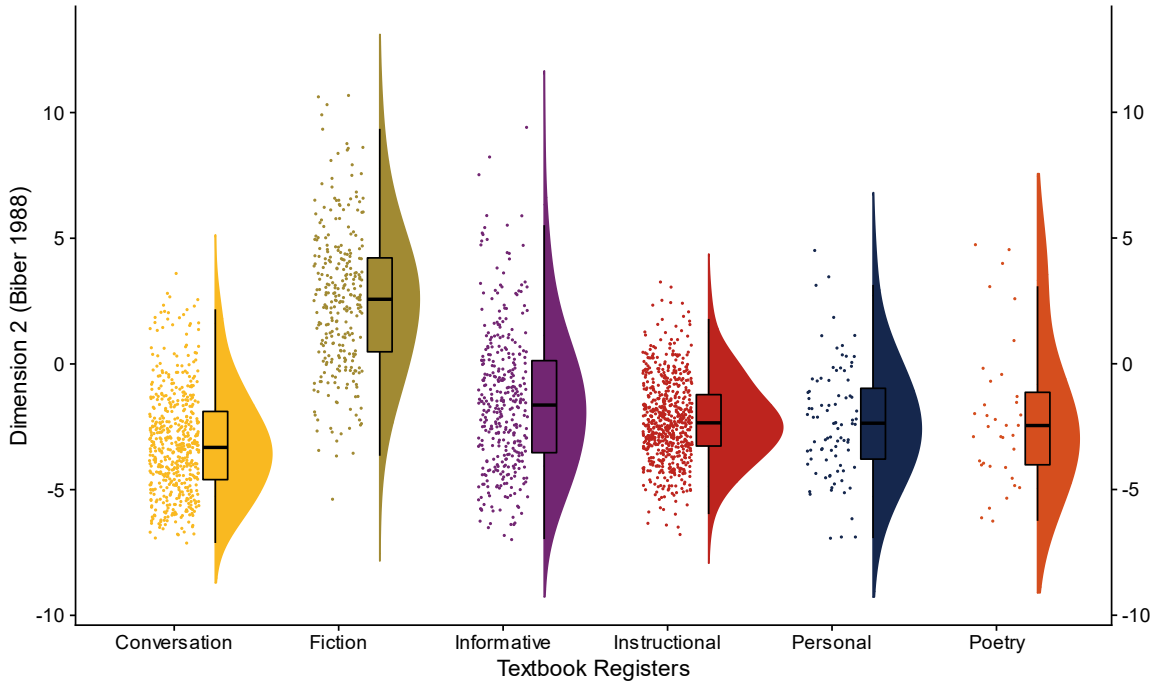


Fig. 33: Distribution of the texts of the TEC on Biber's Dimension 2 (excluding one outlier text from the Poetry register)

As explained in 6.2.6, Poetry & rhyme texts were excluded from all statistical analyses, so that this outlier has not unduly influenced the model summarised in Table 49. Although considerably less intra-textbook register-based variation is observed on Biber's (1988) second dimension than on the first, the analysis of the pair contrasts shows that the differences in Dimension 2 means are significant for all the contrasts involving Conversation and Fiction ($p < .001$). A comparison of the $R^2_{marginal}$ and $R^2_{conditional}$ also makes very clear that the textbook series does not contribute to any significant differences in Dimension 2 scores.

Table 49: Summary of the model: lmer(Dim2 ~ Register + Level + Level*Register + (1|Series))

Predictors	*Estimates*	*CI*	*p-value*
(Intercept) [Conversation, Level A]	-4.09	-4.52 – -3.65	**<0.001**
Register [Fiction]	3.95	3.14 – 4.76	**<0.001**
Register [Informative]	0.71	-0.13 – 1.56	0.097
Register [Instructional]	1.36	0.77 – 1.95	**<0.001**
Register [Personal]	0.57	-0.64 – 1.79	0.357
Level [B]	0.27	-0.30 – 0.83	0.357
Level [C]	0.90	0.33 – 1.46	0.002
Level [D]	1.74	1.17 – 2.31	**<0.001**

Level [E]	1.78	1.14 – 2.41	**<0.001**
*Register [Fiction] * Level [B]*	2.79	1.74 – 3.84	**<0.001**
*Register [Informative] * Level [B]*	0.69	-0.37 – 1.76	0.203
*Register [Instructional] * Level [B]*	-0.19	-0.97 – 0.59	0.629
*Register [Personal] * Level [B]*	-0.21	-1.75 – 1.34	0.790
*Register [Fiction] * Level [C]*	2.71	1.66 – 3.77	**<0.001**
*Register [Informative] * Level [C]*	0.77	-0.25 – 1.79	0.140
*Register [Instructional] * Level [C]*	-0.46	-1.23 – 0.32	0.246
*Register [Personal] * Level [C]*	0.65	-0.92 – 2.22	0.415
*Register [Fiction] * Level [D]*	1.25	0.23 – 2.27	0.016
*Register [Informative] * Level [D]*	0.69	-0.32 – 1.70	0.183
*Register [Instructional] * Level [D]*	-1.05	-1.83 – -0.28	**0.008**
*Register [Personal] * Level [D]*	0.26	-1.37 – 1.90	0.753
*Register [Fiction] * Level [E]*	0.97	-0.09 – 2.02	0.072
*Register [Informative] * Level [E]*	0.84	-0.22 – 1.91	0.120
*Register [Instructional] * Level [E]*	-0.68	-1.53 – 0.16	0.113
*Register [Personal] * Level [E]*	0.67	-0.97 – 2.32	0.423
	Random Effects		
σ^2	4.40		
$\tau_{00\ Series}$	0.01		
ICC	0.00		
$N_{\ Series}$	9		
Observations	1912		
Marginal R^2 / Conditional R^2	0.472 / 0.474		

As illustrated in Fig. 33, the Dimension 2 scores of Textbook Fiction and Informative texts are much more dispersed than those of the other textbook registers. Textbook Fiction and Informative texts at the higher end of the scale tend to feature short sentences with many verbs, the vast majority of which are in the past form and/or perfect aspect, and used with third-person pronouns – three of the six features with positive loadings on Biber's narrative dimension, e.g., (183)–(184). In terms of semantics, many of these verbs belong to Biber's public verb category (e.g., DISCUSS, SAY, WRITE). In addition, we find higher-than-average occurrences of present participial clauses and synthetic negation (e.g., *her mother had had no idea*), e.g., (183)–(184).

(183) Cal **climbed** over the sheet of rusted metal, **sniffing** the air. **He could** detect ammonia and sulphur. But there **was** also something else, something special. **He called** out to **his** sister and immediately **she was** there, **jumping** down from a blackened metal beam overhead. **Her** green eyes **stared** at **him** out of **her** scarred and blistered face.
'Keep watch.'
She nodded and **turned**, **looking** out over a nightmarish landscape of twisted metal and smashed vehicles. Cal **dropped** to all fours and **crawled** towards a hole in the metal canopy. **He leaned** over the edge and **peered** into the darkness. First **he had smelt** them, now **he could** see **them**: pale

blue shapes **glimmering** in the darkness. **They were** mushrooms. **They were** food! <TEC: Achievers B1+>

(184) The very first cell phone novel **was written** in 2003 by a man in Tokyo who called **himself** Yoshi. It **became** so popular, mainly through word of mouth, that it **was** later **published** as a paperback. The book version **sold** 2.6 million copies and a television series, a comic book and a film **were made** of the story. [...] Although the idea **originated** in Japan, cell phone novels have also **sprung** up in the rest of East Asia, Europe and Africa. Many are **written** by high school or university students who are very familiar with the topics that teenagers are interested in. Twenty-one-year-old Rin **said** that **she started her** novel during **her** final year at high school and **explained** that it **was** the tragic love story of two childhood friends. [...] Rin said that her mother **had had** no idea that she **had been writing** a novel and **was** therefore very surprised when **she saw** a book with **her** daughter's name on it. <TEC: Solutions Intermediate>

Since only linguistic features with positive loadings contribute to Biber's (1988) 'Narrative vs. Non-Narrative Concerns' dimension (see Table 40), at the lower end of the Dimension 2 scale, we find texts with particularly low counts of the features mentioned above. The majority of these texts were extracted from beginner and lower-intermediate textbooks, thus largely before the past tense and perfect aspect are introduced to English learners. It is worth noting that many of the features of Dimension 2 can be considered opposites to those with positive loadings on Dimension 1: present vs. past tense, first- and second-person vs. third-person pronouns, and private vs. public verbs. Thus, it comes as no surprise that Textbook Conversation and Informative texts with high scores on Dimension 1 also tend to score low on Dimension 2. For instance, the text from which example (185) was extracted scores 21.02 on Dimension 1 and -6.76 on Dimension 2. In extract (185), features with positive loadings on Dimension 1 are highlighted. In this extract, there are zero occurrences of any of the six features that load on Dimension 2.

(185) Lucy: **Hi**, Sam, Justin. **What's up**?
Sam: **Hi** Lucy, **hi** Maya. Look, Justin **has** a video camera and **we want** to make a film.
Justin: **What ... we want** to make a film?
Lucy: T**hat's** a good idea! But **what's** the film **about**?
Sam: **It's** an action film. A kung fu film! **Do you** want to help?
Justin: An action film? Kung fu? But ...
Lucy: Sorry, Sam, but **I don't like** action films.
Sam: **Well, do you** have any other ideas?
Justin: Sorry, but **I** want to make a video about Plymouth for **my** dad.
<TEC: Access G 1>

The model summary in Table 49 also highlights a number of significant interactions between three textbook registers and three textbook proficiency levels. Since the magnitude of these effects is not immediately obvious from the model summary, the Dimension 2 estimates are plotted per register and level in Fig. 34. The mean Dimension 2 estimates are also listed in

Table 50. The scatterplots in Fig. 34 make clear that, as expected, for most registers, beginner textbooks feature very few of the features that constitute Dimension 2, especially those that have yet to be introduced to learners at this early stage of language acquisition, e.g., the past tense, perfect aspect, and present participial clauses. The difference across the different proficiency levels is greatest for the Fiction register, which reaches a peak in Dimension 2 scores in the fictional texts of level C textbooks, corresponding with the year in which the textbooks series tend to have a strong focus on the past tense and the perfect aspect, and then does not see any significant changes across the other two levels. Personal correspondence, Conversation and Informative texts all see a gradual increase in Dimension 2 scores with increasing proficiency levels, though as reported in Table 49, many of these differences are not significant and, indeed,
Table 50 shows many overlapping confidence intervals across textbook levels.

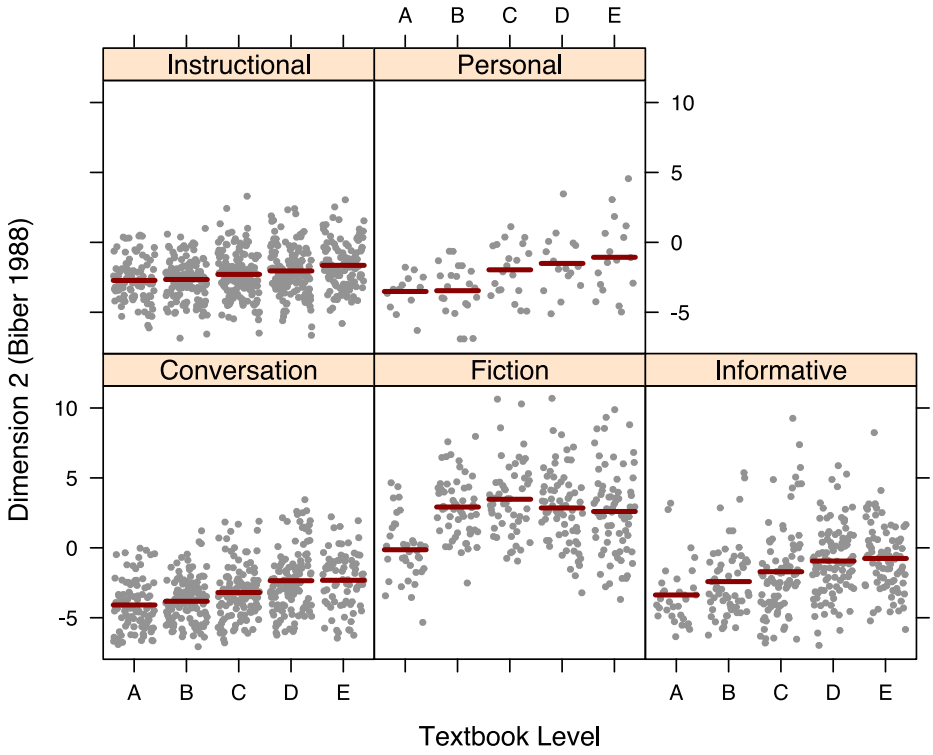


Fig. 34: Visualisation of the effects of Register*Level interactions on the estimated Dimension 2 values of the TEC texts (as calculated with the model summarised in Table 49)

Table 50: Estimated Dimension 2 means and 95% confidence intervals (CI)

Register	Level	mean	SE	df	lower.CL	upper.CL
Conversation	*A*	-4.09	0.22	870	-4.52	-3.65
Conversation	*B*	-3.82	0.20	680	-4.21	-3.43
Conversation	*C*	-3.19	0.20	633	-3.58	-2.80
Conversation	*D*	-2.34	0.20	626	-2.74	-1.95
Conversation	*E*	-2.31	0.25	982	-2.80	-1.82
Fiction	*A*	-0.14	0.36	1547	-0.84	0.56
Fiction	*B*	2.92	0.29	1172	2.35	3.49
Fiction	*C*	3.48	0.30	1341	2.90	4.05
Fiction	*D*	2.86	0.26	1154	2.35	3.37
Fiction	*E*	2.61	0.25	1116	2.11	3.10
Informative	*A*	-3.37	0.38	1722	-4.11	-2.63
Informative	*B*	-2.41	0.28	1335	-2.96	-1.87
Informative	*C*	-1.71	0.23	927	-2.16	-1.25
Informative	*D*	-0.94	0.22	893	-1.37	-0.51
Informative	*E*	-0.75	0.23	954	-1.21	-0.30
Instructional	*A*	-2.72	0.21	899	-3.14	-2.31
Instructional	*B*	-2.65	0.19	649	-3.02	-2.28
Instructional	*C*	-2.28	0.18	589	-2.64	-1.93
Instructional	*D*	-2.03	0.18	580	-2.39	-1.68
Instructional	*E*	-1.63	0.20	660	-2.02	-1.24
Personal correspondence	*A*	-3.51	0.59	1918	-4.67	-2.36
Personal correspondence	*B*	-3.46	0.45	1841	-4.35	-2.57
Personal correspondence	*C*	-1.96	0.47	1888	-2.89	-1.03
Personal correspondence	*D*	-1.51	0.53	1906	-2.55	-0.47
Personal correspondence	*E*	-1.07	0.51	1890	-2.08	-0.06

6.3.1.3 Intra-textbook variation on Biber's Dimension 3

On Biber's (1988) third dimension, textbook registers exhibit few marked differences in their scores. The distribution of Dimension 3 scores on Fig. 35 and comparisons for relative contrast effects suggest two clusters of textbook registers on this dimension. On the upper end of the range, we find the Instructional ($\bar{x}$ = 3.52, SD = 2.82) and Informative texts ($\bar{x}$ = 3.12, SD = 3.07) and, on the lower end, Personal correspondence ($\bar{x}$ = -0.83, SD = 2.96), Conversation ($\bar{x}$ = -0.45, SD = 2.37), and Fiction texts ($\bar{x}$ = -0.65, SD = 2.74). The differences in scores are significant across the clusters ($p < .001$), but there are no significant differences between the scores of the registers within each cluster.

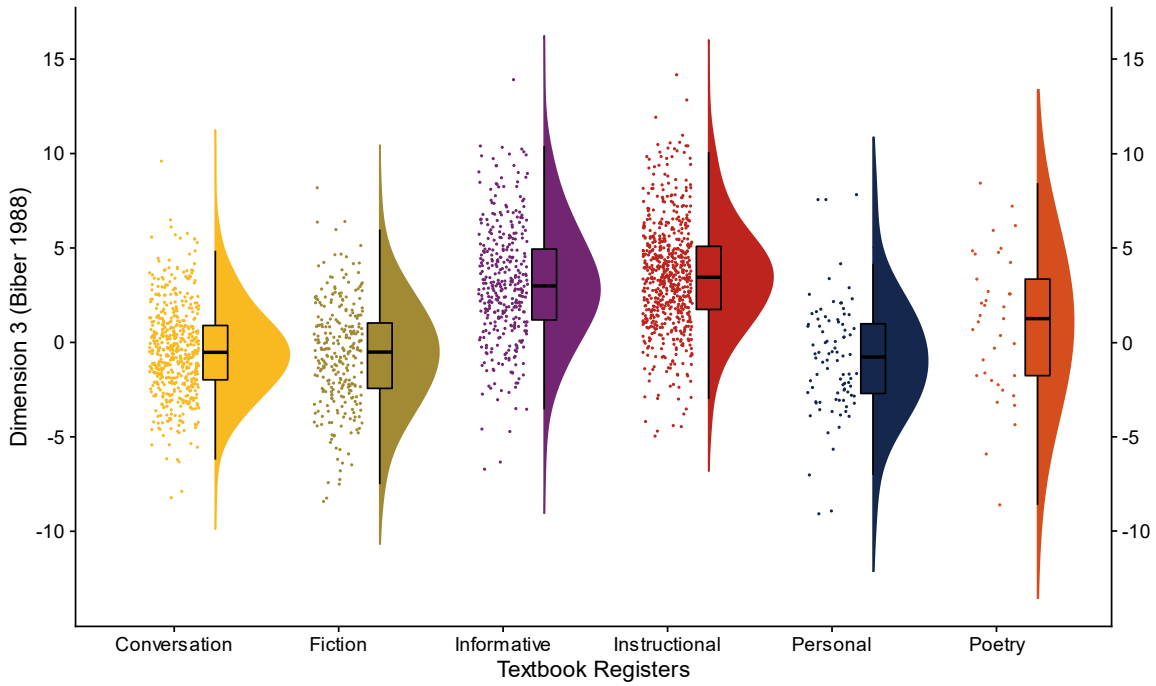


Fig. 35: Distribution of the texts of the TEC on Biber's (1988) Dimension 3

Table 51: Summary of the model: lmer(Dim3 ~ Register + Level + Level*Register + (1|Series))

Predictors	*Estimates*	*CI*	*p-value*
(Intercept) [Conversation, Level A]	-0.18	-1.13 – 0.77	0.715
Register [Fiction]	-0.15	-1.10 – 0.81	0.764
Register [Informative]	3.03	2.04 – 4.03	**<0.001**
Register [Instructional]	5.16	4.46 – 5.85	**<0.001**
Register [Personal]	-0.04	-1.47 – 1.39	0.957
Level [B]	-0.29	-0.96 – 0.37	0.389
Level [C]	-0.24	-0.91 – 0.42	0.471
Level [D]	0.52	-0.15 – 1.19	0.128
Level [E]	-0.08	-0.84 – 0.67	0.829
*Register [Fiction] * Level [B]*	0.35	-0.89 – 1.58	0.582
*Register [Informative] * Level [B]*	0.48	-0.77 – 1.74	0.45
*Register [Instructional] * Level [B]*	-0.89	-1.81 – 0.03	0.058
*Register [Personal] * Level [B]*	0.18	-1.64 – 2.00	0.844
*Register [Fiction] * Level [C]*	-0.15	-1.39 – 1.09	0.809
*Register [Informative] * Level [C]*	1.08	-0.12 – 2.28	0.077
*Register [Instructional] * Level [C]*	-1.52	-2.43 – -0.61	**0.001**
*Register [Personal] * Level [C]*	-0.56	-2.41 – 1.28	0.55
*Register [Fiction] * Level [D]*	-0.05	-1.25 – 1.15	0.934
*Register [Informative] * Level [D]*	-0.14	-1.33 – 1.06	0.821
*Register [Instructional] * Level [D]*	-2.13	-3.04 – -1.21	**<0.001**
*Register [Personal] * Level [D]*	-2.10	-4.03 – -0.17	**0.033**
*Register [Fiction] * Level [E]*	-0.52	-1.76 – 0.72	0.411
*Register [Informative] * Level [E]*	1.07	-0.18 – 2.32	0.094
*Register [Instructional] * Level [E]*	-1.08	-2.07 – -0.08	**0.033**

*Register [Personal] * Level [E]*	0.18	-1.75 – 2.12	0.852
	Random Effects		
σ^2	6.08		
$\tau_{00\ Series}$	1.51		
ICC	0.2		
$N_{\ Series}$	9		
Observations	1912		
Marginal R^2 / Conditional R^2	0.353 / 0.482		

According to Biber's interpretation of this third dimension, textbook registers are characteristically situation-dependent, thus more like broadcasts and conversation, rather than explicit, as in official documents and professional letters. However, it should be noted that most of the features with positive loadings on this dimension, WH-relative clauses, pied-piping constructions, and nominalisations, are comparatively rare across all textbook registers. That said, these are fairly rare features in general English and the textbook texts being relatively short, the median count for WH-relative clauses on object and subject positions and pied-piping constructions is zero, thus very low *z*-scores are observed across all registers for this feature. According to the counts generated by the MAT (see 6.2.3), nominalisations are more frequent in instructional language than in any other textbook register. This is due to how nominalisations are operationalised in Biber's original MDA: all noun lemmas ending in *-tion, -ment, -ness,* and *-ity* are counted as instances of nominalisations (Biber 1988: 227). As a result, this variable is highly inflated by a number of lemmas that are very frequent in textbooks' instructions and explanations: foremost *activity,* followed by *statement, document, comment*, *argument* and *element* (186)–(188), even though many of these nouns are not necessarily foremost derived from a verb, e.g., *document* and *element*.

(186) In pairs, read out your **statements** and guess if your partner's **statements** are true or false. <TEC: Achievers A1>

(187) Team **activity**: Impressions of New York
a) In groups of four, write each of these **statements** about New York on a piece of paper: <TEC: Green Line 4>

(188) What sport is the **document** about? b. Find important qualities related to this **activity**. <TEC: Hi There 5[e]>

For both the Instructional and Informative registers, the variables which contribute most to their high scores on Dimension 3 are phrasal coordination, with high mean *z*-scores (which have a positive loading on this dimension), as well as adverbs (with a negative loading), for which the corresponding *z*-scores are very low. For the Biber Tagger, and correspondingly also for the MAT, phrasal coordination refers to two adjectives, adverbs, verbs, or nouns separated by *and*. This is a pattern which occurs more frequently across all textbook registers than on average in Biber's (1988) general English corpus, but which is particularly frequent in instructional and informative

texts of the TEC. Verb + *and* + verb trigrams are particularly frequent in textbook instructions, e.g., (189)–(190). Though it is worth noting that the lemma list of these verb combinations spans more than 300 different combinations, the most frequent ones are: *ask and answer* ($n = 283$), *listen and check* ($n = 251$), *read and listen* ($n = 76$), *listen and repeat* ($n = 66$), *check and repeat* ($n = 53$), *listen and write* ($n = 25$), *listen and say* ($n = 25$), *listen and find* ($n = 21$), *check and correct* ($n = 18$), *look and speak* ($n = 17$), *listen and match* ($n = 15$), *listen and answer* ($n = 15$), *like and do* ($n = 15$), *listen and complete* ($n = 13$), *stop and think* ($n = 11$), *listen and identify* ($n = 11$) and *read and complete* ($n = 10$).

(189) Pronunciation (Different stress in **German and English** words)
a) Write down the **German and English** words. Underline the stressed vowel in each word.
b) **Listen and check**. **True and false** friends. <TEC: Access G 3>

(190) Jamie's school dinners. a. **Read and learn** about Jamie's campaign. <TEC: Hi There 6e>

Although noun + *and* + noun combinations are the most frequent type of phrasal coordination in both registers, these are considerably less formulaic than the verb + *and* + verb collocations listed above. In instructions, the most frequent nominal phrasal coordination trigrams are *words and phrases* ($n = 121$), *words and expressions* ($n = 55$), *A and B* ($n = 49$), *question(s) and answer(s)* ($n = 47$), *pros and cons* ($n = 27$), and *advantages and disadvantages* ($n = 25$). Adjective + *and* + adjective and adverb + *and* + adverb trigrams are more frequent in informative texts than in instructions but, apart from the trigrams *more and more* ($n = 29$) and *again and again* ($n = 8$), these also display considerably less formulaicity and thus the Textbook Informative subcorpus features almost as many trigram types as tokens, e.g., (192)–(193). In the Instructional register, the most frequent adjectival trigrams are *positive and negative* ($n = 21$) and *good and bad* ($n = 12$), as well as several phrases specific to language learning, e.g., (194)–(195).

(191) TV is so 20th century! **More and more** young people are using online social networks like my Space and facebook.com, popular in many different countries. <TEC: English in Mind 3>

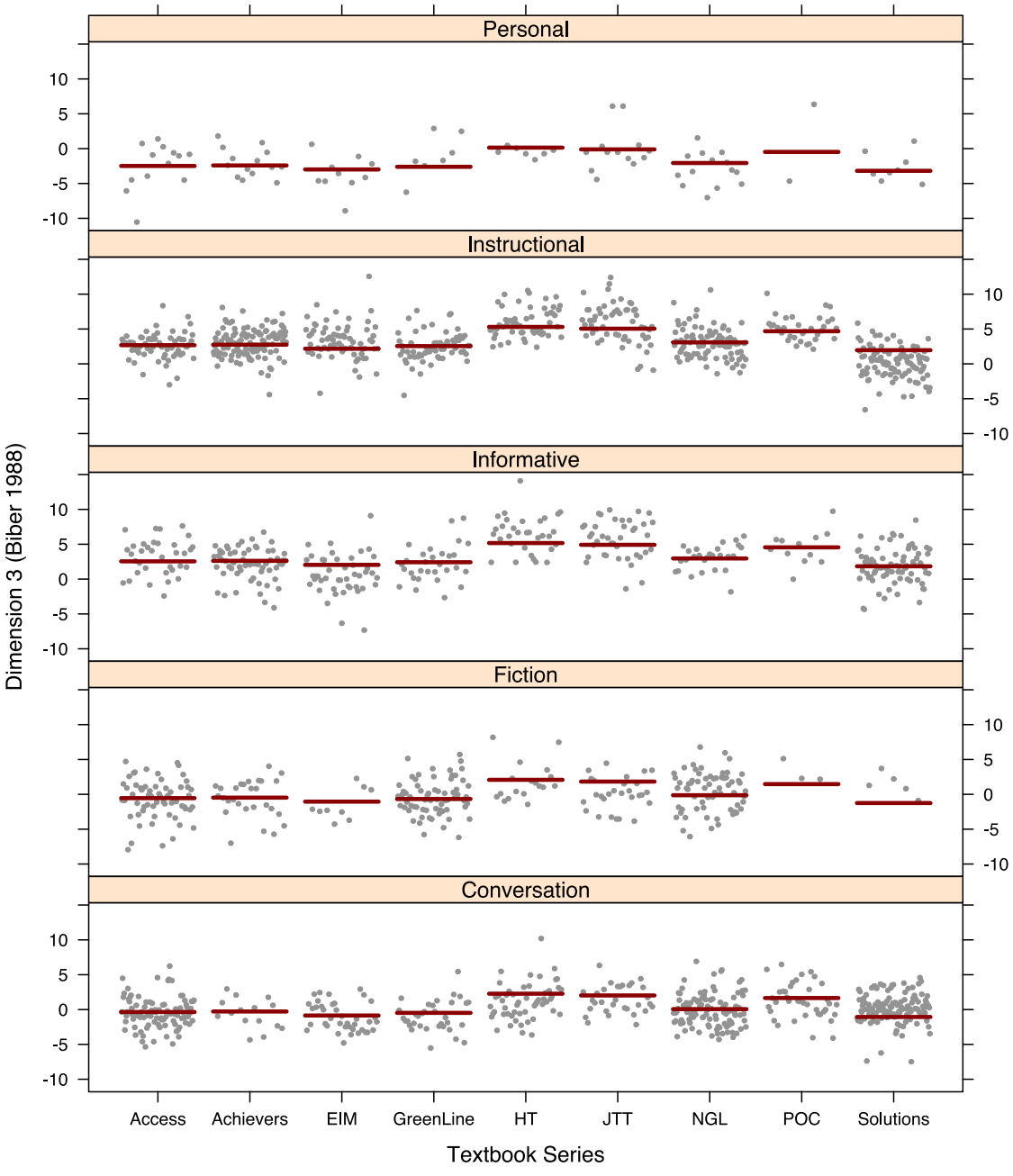
(192) The fresh sea air and the warm sun are nice. But the beaches are not always warm and sandy. Sometimes they are cold, **windy and rocky**. <TEC: Green Line 1>

(193) Gumbo is a dish from Louisiana, a kind of stew with many ingredients. There are different ways of making it, but it usually has vegetables, meat, fish, **herbs and spices** in it. […] It is eaten by **young and old**, **black and white**, **rich and poor**, both every day and on important occasions. <TEC: Access G 4>

(194) Finish the dialogue. Use **positive and negative** verbs. <TEC: Access G 1>

(195) YOU MUST: use the past tense (**regular and irregular** verbs), use negatives. <TEC: Piece of Cake 5e>

As compared with the previous two dimensions, Dimension 3 scores display the greatest influence due to textbook series-based idiosyncrasies: the mixed-effects model that predicts Biber's (1988) Dimension 3 scores without the by-series random effect structure explains 35% of the total variance among the texts of the TEC (see $R^2_{marginal}$ on Table 51) whilst, by adding by-series intercepts, its predictive power reaches 48% (see $R^2_{marginal}$ on Table 51). This effect is visualised in Fig. 36



on which the three French textbook series, *Hi There* (HT), *Join the Team* (JTT) and *Piece of Cake* (POC), stand out as having markedly higher Dimension 3 scores across all registers. Two sets of features contribute to this effect. First, phrasal coordination, which, as discussed above, is generally very prominent in Textbook English, appears to be even more frequent in French textbooks. Second, three adverbial features which contribute negatively to Dimension 3 scores, time adverbials, place adverbials and

general adverbs[44] are comparatively rare in these three French series as compared to the German and Spanish series of the TEC.

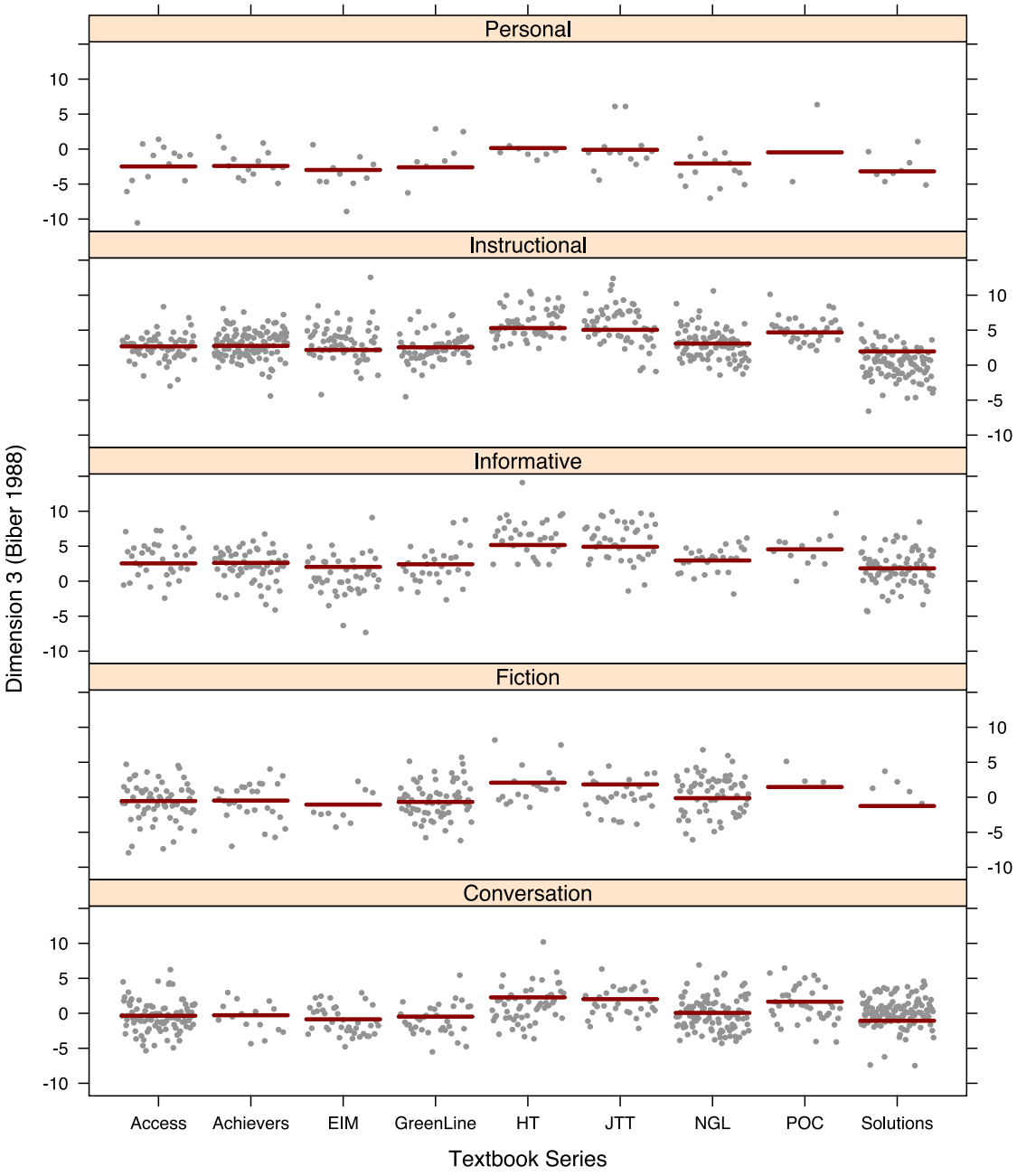


Fig. 36: Estimated scores on Biber's (1988) Dimension 3 subdivided by register and textbook series

[44] Biber's 'general adverbs' is a broad category that includes prototypical adverbs such as *easily, remarkably, yet* and *never*, question words, the particles of some phrasal verbs such as *go back* and *turn around* and, following Biber's [1988] tagging rules, also encompasses many words that frequently function as discourse markers, e.g., *so*, *right, yes, anyway,* whilst excluding adverbs counted in more specific categories such as time and place adverbials, hedges, amplifiers and downtoners. This general adverbs category is problematic both linguistically, because it somewhat arbitrarily agglomerates adverbs with very different functions, and, technically, since the MAT is known to inflate *z*-scores for this category (Nini 2014: 13) – the cause of which has yet to be identified (Nini 2019: 77 and personal communication).

6.3.1.4 Intra-textbook variation on Biber's Dimension 4

Fig. 37 shows that the mean textbook register scores on Biber's (1988) fourth dimension are all negative. The original raincloud plot revealed an outlier in the Informative register which was removed from the plot printed below. It corresponds to a text from *Join the Team 5*[e] which gives advice about how "princesses" should respond if "a guy" asks them out on a date, an extract of which is printed below (196). Its primary function is to persuade, so it comes as no surprise that it should score so high on this 'Overt Persuasion' dimension. Indeed, it features singularly high normalised counts for five of the six features that make positive contributions to this dimension: infinitives (e.g., *go, prepare*), predictive modals (*will, would*), suasive verbs (e.g., *ASK*, *ALLOW*), conditional subordination (*if*) and necessity modals (*must*).

(196) If HE **asks** YOU out...
You lucky girl! He **asked**! He finally **asked**! [...] Be enthusiastic, but be cool.
If you are like me, and your father, the prince of a small European country, **won't allow** you to **go** out with a boy he hasn't met, you **must** confess this IMMEDIATELY to any boy who **asks** you out. [...] He needs time to **prepare** mentally [...].
If you have to **check** with your parents before accepting a date, say « Oh, I'**d** love to **go** to the planetarium with you on Saturday, but I have to check with my mom first. May I call you back when I know for sure?
Then be sure to **call** him back promptly.
<TEC: Join the Team 5e, ellipses in the original>

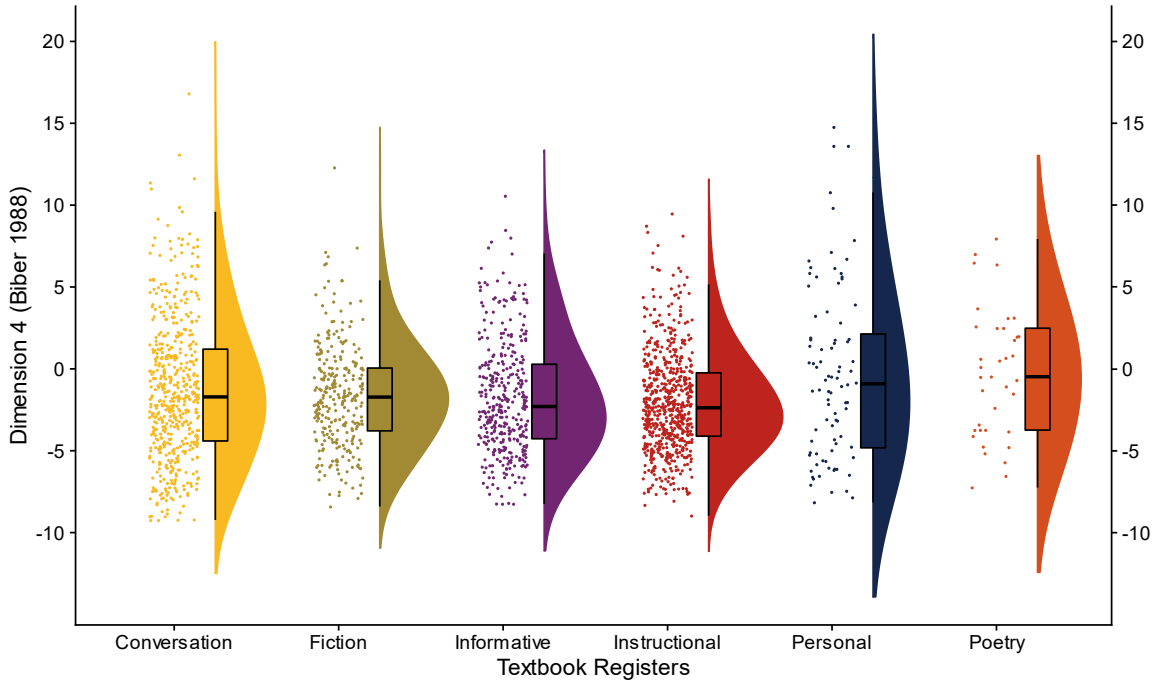


Fig. 37: Distribution of the texts of the TEC on Biber's (1988) Dimension 4 (with one outlier from the Informative register removed)

In order for it to not unduly influence the model, this outlier text was also removed from the dataset used to model Dimension 4 scores. As with Dimensions 2 and 3, this 'Overt Persuasion' dimension is also susceptible to significant Register*Level effects (see Table 52). These are visualised in Fig. 38, which shows that all six linguistic features that load on Dimension 4 are least frequent in Level A textbooks, and slightly more frequent, though still less than average, in Level B textbooks. The differences between the three more advanced levels are not significant. Whilst it was to be expected that *if*-sentences, and consequently conditional subordinators (*if*, *unless*) and, to a lesser extent, prediction modals (*will*, *would* and *shall*), as well as split auxiliaries are not introduced until more advanced textbooks, the most notable difference between beginner and the more advanced textbooks is found in the comparatively low relative frequencies of *to*-infinitives in the lower-level textbooks. The most frequent lexico-grammatical patterns associated with *to*-infinitives, *WANT to, BE going to, HAVE to, LIKE to, TRY to,* are largely the same across the five textbook levels but they are much more frequent in the pedagogical materials designed for EFL students in their third, fourth and fifth year of secondary school than in the first two years of tuition. It is also worth bearing in mind that some textbook authors do not introduce the *BE going to* future construction in the first volume of their studies (e.g., *Access G* and *English in Mind*) – a choice that will inevitably also influence these Dimension 3 scores.

Table 52: Summary of the model: lmer(Dim4 ~ Register + Level + Level*Register + (1|Series)) (excluding the outlier discussed above)

Predictors	*Estimates*	*CI*	*p-value*
(Intercept) [Conversation, Level A]	-5.74	-6.56 – -4.91	**<0.001**
Register [Fiction]	1.01	-0.23 – 2.24	0.11
Register [Informative]	0.98	-0.30 – 2.27	0.134
Register [Instructional]	1.75	0.85 – 2.64	**<0.001**
Register [Personal]	0.49	-1.37 – 2.34	0.608
Level [B]	3.83	2.97 – 4.70	**<0.001**
Level [C]	5.96	5.11 – 6.82	**<0.001**
Level [D]	5.69	4.82 – 6.56	**<0.001**
Level [E]	6.38	5.39 – 7.36	**<0.001**
*Register [Fiction] * Level [B]*	-1.59	-3.19 – 0.00	0.05
*Register [Informative] * Level [B]*	-2.43	-4.06 – -0.80	**0.003**
*Register [Instructional] * Level [B]*	-2.49	-3.67 – -1.30	**<0.001**
*Register [Personal] * Level [B]*	-0.11	-2.46 – 2.24	0.928
*Register [Fiction] * Level [C]*	-1.82	-3.42 – -0.22	**0.026**
*Register [Informative] * Level [C]*	-2.89	-4.44 – -1.33	**<0.001**
*Register [Instructional] * Level [C]*	-3.94	-5.12 – -2.76	**<0.001**
*Register [Personal] * Level [C]*	-0.32	-2.71 – 2.07	0.793
*Register [Fiction] * Level [D]*	-1.21	-2.76 – 0.35	0.128
*Register [Informative] * Level [D]*	-2.13	-3.67 – -0.59	**0.007**
*Register [Instructional] * Level [D]*	-3.59	-4.77 – -2.40	**<0.001**

*Register [Personal] * Level [D]*	2.76	0.27 – 5.26	**0.030**
*Register [Fiction] * Level [E]*	-2.2	-3.81 – -0.59	**0.007**
*Register [Informative] * Level [E]*	-2.39	-4.02 – -0.77	**0.004**
*Register [Instructional] * Level [E]*	-2.85	-4.14 – -1.56	**<0.001**
*Register [Personal] * Level [E]*	-0.12	-2.63 – 2.38	0.923
	Random Effects		
σ^2	10.18		
$\tau_{00\ Series}$	0.59		
ICC	0.06		
$N_{\ Series}$	9		
Observations	1911		
Marginal R^2 / Conditional R^2	0.213 / 0.256		

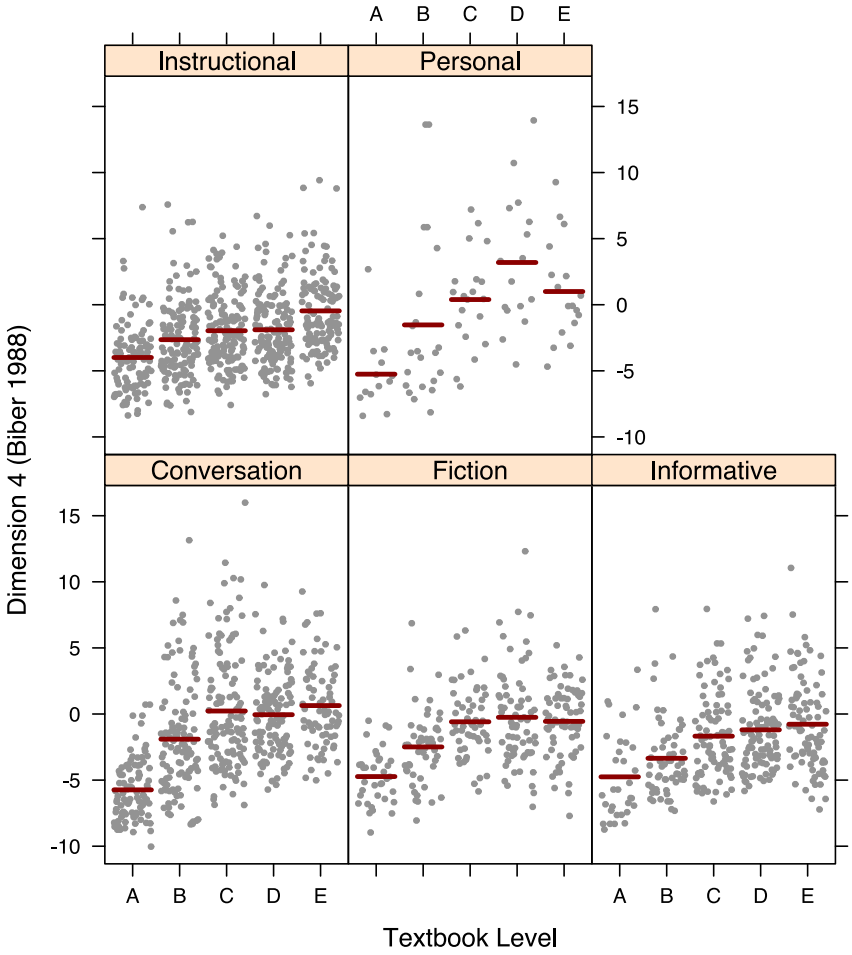


Fig. 38: Estimated mean Dimension 4 scores subdivided per textbook register and proficiency level

6.3.1.5 Intra-textbook variation on Biber's Dimension 5

The scores on Biber's (1988) fifth and sixth dimensions are only marginally useful for the exploration of Textbook English registers because they are made up of just six and four positive-loading features respectively (see Table 40) – all of which are relatively rare in Textbook English (see Table 43). As a result, for many textbook

texts, the normalised counts for these features are equal to zero. This leads to floor effects, as seen in the corresponding rainplots (see Fig. 39 and 40). For Dimension 5, the floor level is -3.92: the score obtained for 265 out of the 1949 textbook texts processed and analysed for this additive MDA (14%) (see Fig. 39).

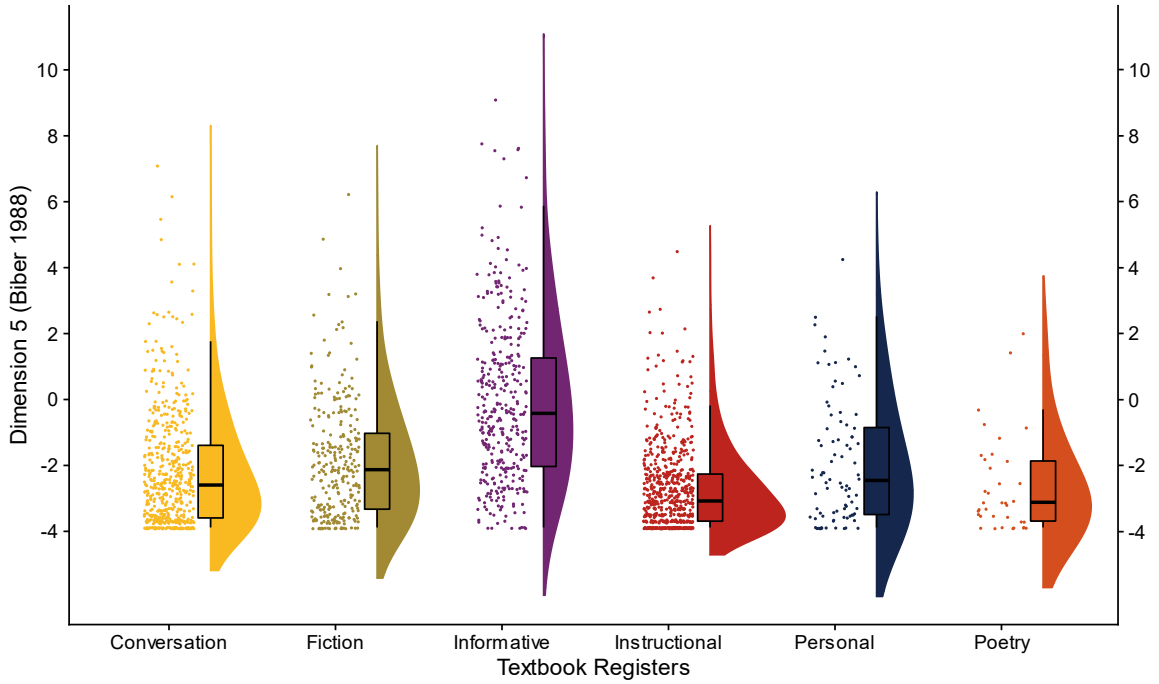


Fig. 39: Distribution of the texts of the TEC on Biber's (1988) Dimension 5

The mean textbook register scores on Biber's fifth 'Abstract vs. Non-Abstract Information' dimension are all negative. This corresponds to Biber's (1988) results since only academic prose and official documents – two registers largely absent from school EFL textbooks – registered positive scores on this dimension (around 5), whilst newspaper writing and popular lore were situated around zero and all other registers scored negatively. Newspaper writing is the closest register to the Informative textbook register and this is the only register with a significantly different mean Dimension 5 score from all other textbook registers ($\bar{x}$ = -0.11, SD = 2.43). However, it should also be noted that the features that make positive contributions to Dimension 5 (conjuncts, agentless passives, past participial clauses, *by*-passives, past participial WHIZ deletion relatives[45], other adverbial subordinators) are not typically introduced until after the first few years of EFL tuition. Indeed, the Dimension 5 scores reported in Table 53 show that, as students are gradually introduced to these features, the scores of the Textbook Informative texts progressively rise from negative values to above zero. It is worth noting that the Level E mean score for Textbook Informative writing roughly corresponds to Biber's (1988) Dimension 5 scores for his

[45] In the TEC, this feature corresponds to constructions such as "Match two words to make collocations **used** in the DVD clips" <TEC: Solutions Intermediate Plus> and "[P]eople sing Auld Lang Syne (The Good Old Days), a song **written** by Robert Burns in the 1780s to honour old and new friends." <TEC: Hi There 5e>.

press reportage, editorials, and reviews subcorpora, though the increasingly large standard deviation values also show that the informative texts featured in the more advanced textbooks vary substantially in their use of the six features that load on Biber's (1988) Dimension 5 (see Table 40).

Table 53: Mean Dimension 5 scores for Informative texts grouped by textbook level

Textbook Level	Mean score (SD)
A	-2.52 (0.99)
B	-1.18 (1.94)
C	0.08 (2.39)
D	0.46 (2.28)
E	0.68 (2.54)

Since the few features that load on Dimensions 5 and 6 are largely absent in beginner textbooks, these two dimensions were only modelled for the texts of intermediate to advanced textbooks (Levels C to E). For Dimension 5, this filtering of the textbook data resulted in just 16 texts scoring the minimum dimension score corresponding to zero occurrences of all six features. However, in spite of this filtering, the distributions of the Dimension 5 and 6 scores entered in the models remain skewed. This leads to violations of the normal distribution of residuals so that the results of the linear mixed-effects models summarised in Tables 54 and 55, in particular the reported *p*-values and confidence intervals, must be interpreted with caution. For both these models, the reference levels are the Conversation for the Register variable and Level C for the proficiency level variable.

Table 54: Summary of the model: lmer(Dim5 ~ Register + Level + Level*Register + (1|Series)) for Levels C to E textbook texts only

Predictors	*Estimates*	*CI*	*p-value*
(Intercept) [Conversation, Level C]	-2.07	-2.58 – -1.55	**<0.001**
Register [Fiction]	0.57	-0.05 – 1.20	0.071
Register [Informative]	1.78	1.25 – 2.31	**<0.001**
Register [Personal]	1.02	0.12 – 1.92	**0.026**
Level [D]	0.51	0.03 – 0.99	**0.036**
Level [E]	0.48	-0.09 – 1.04	0.096
*Register [Fiction] * Level [D]*	0.13	-0.70 – 0.97	0.752
*Register [Informative] * Level [D]*	0.05	-0.68 – 0.78	0.886
*Register [Personal] * Level [D]*	-0.96	-2.29 – 0.38	0.160
*Register [Fiction] * Level [E]*	-0.39	-1.26 – 0.48	0.375
*Register [Informative] * Level [E]*	0.67	-0.12 – 1.46	0.095
*Register [Personal] * Level [E]*	-0.62	-1.97 – 0.72	0.363
	Random Effects		
σ^2	3.57		
$\tau_{00\ Series}$	0.34		

ICC	0.09
N_{Series}	9
Observations	833
Marginal R^2 / *Conditional* R^2	0.184 / 0.255

The model designed to predict the scores of the intermediate and advanced textbooks on Biber's fifth 'Abstract vs. Non-Abstract Information' dimension explains just 25% of the variance (see Table 54). As also illustrated in Fig. 39, relatively little register-based variation can be observed on this dimension. The main discernible effect consists in the higher mean score for Informative texts, which, following Biber's interpretation of Dimension 5, may be considered more abstract than the other four textbook registers examined.

6.3.1.6 Intra-textbook variation on Biber's Dimension 6

Biber's (1988) sixth dimension is calculated on the basis of just four features: *that*-clauses as verb complements, *that*-relative clauses on object position, *that*-clauses as adjective complements and demonstratives, all which make positive contributions to this dimension. Since the first three of these are relatively low frequency features in general English, many textbook texts have a count of zero, thus the corresponding rainplot (see Fig. 40) features overwhelmingly negative scores: 66/1949 texts (3%) have the lowest value possible of -3.495 which corresponds to an absence of all four features in these texts. As hinted at in Fig. 40, these minimum scores are fairly evenly distributed across the six registers.

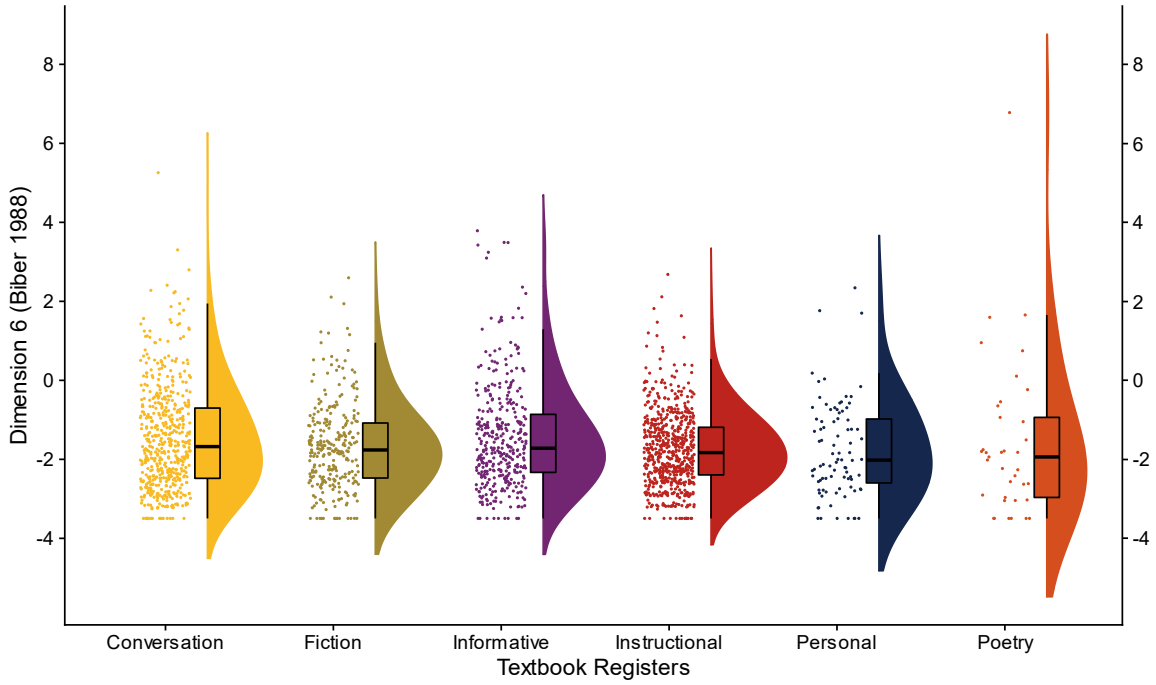


Fig. 40: Distribution of the texts of the TEC on Biber's (1988) Dimension 6

Fig. 40 also features two outliers in the Conversation and Poetry registers. The former corresponds to a dialogue set during a guided tour of a touristic site (197). As all of the interlocutors are in the same environment, demonstrative pronouns are frequently employed and the tour guide's reporting of the legend results in an unusually high density of *that* clauses as verbal complements of verbs such as *BELIEVE*, *THINK* and *SEE*.

(197) OK now we are in the centre of Tintagel Castle. **This** part is called the Island Courtyard. It was once a great hall for celebrations and feasts and meetings.
Now let's go through **this** gate. Is everybody here?
OK, here's a question for all of you: Why did Prince Richard build a castle here? [...]
Maybe **this** is what he thought: "Many people here believe **that** King Arthur lived at Tintagel. If I build a castle here, maybe they will think **that** it was once King Arthur's castle. Then if I live in this castle, they will think **that** King Arthur and I are from the same family.
Yeah, but couldn't people see **that** it was a new castle, and not King Arthur's castle? <TEC: Access G 2>

The second outlier corresponds to the poem *The Lost Generation* by Jonathan Reed (198), which can be read line-by-line both from top to bottom and from bottom to top. In order to achieve this, the poet makes extensive use of *that*-clauses as verb complements, which results in the unusually high score of nearly 6 on Biber's 'On-Line Informational Elaboration' dimension.

(198) I am part of a lost generation
and I refuse to believe **that**
I can change the world
I realize this may be a shock but
"Happiness comes from within"
is a lie, and
"Money will make me happy."
So in 30 years I will tell my children
they are not the most important thing in my life
My employer will know **that**
I have my priorities straight because
work
is more important than
family
I tell you this
I will live in a country of my own making
In the future
Environmental destruction will be the norm
No longer can it be said **that**
My peers and I care about this earth
It will be evident **that**
My generation is apathetic and lethargic
It is foolish to presume **that**
There is hope. <TEC: Solutions Intermediate>

Aside from these two outliers, however, no noticeable register differences on Biber's (1988) Dimension 6 can be discerned. This is not surprising given the generally very low frequencies of the four features that load on this dimension and is confirmed by the model summarised in Table 55, which, as explained in 6.3.1.5, only covers textbook levels C to E. As compared to Biber's (1988) first five dimensions, the model for Dimension 6 explains the least amount of variance (11%).

Table 55: Summary of the model: lmer(Dim6 ~ Register + Level + Level*Register + (1|Series)) for levels C to E textbooks only

Predictors	*Estimates*	*CI*	*p-value*
(Intercept) [Conversation, Level C]	-1.67	-1.98 – -1.36	**<0.001**
Register [Fiction]	0.39	-0.01 – 0.78	0.055
Register [Informative]	0.17	-0.16 – 0.51	0.308
Register [Personal]	-0.05	-0.62 – 0.53	0.876
Level [D]	0.52	0.21 – 0.82	**0.001**
Level [E]	0.94	0.58 – 1.29	**<0.001**
*Register [Fiction] * Level [D]*	-0.67	-1.20 – -0.14	**0.013**
*Register [Informative] * Level [D]*	-0.27	-0.74 – 0.19	0.248
*Register [Personal] * Level [D]*	-0.35	-1.20 – 0.50	0.416
*Register [Fiction] * Level [E]*	-1.10	-1.65 – -0.54	**<0.001**
*Register [Informative] * Level [E]*	-0.82	-1.32 – -0.32	**0.001**
*Register [Personal] * Level [E]*	-0.14	-0.99 – 0.71	0.747
	Random Effects		
σ^2	1.44		
$\tau_{00\ Series}$	0.11		
ICC	0.07		
$N_{\ Series}$	9		
Observations	833		
Marginal R^2 / Conditional R^2	0.042 / 0.112		

In spite of these very low R^2 values and the relatively small differences in the coefficients listed in Table 55, the model summary nevertheless invites further exploration of the Register*Level interactions, which the model returns as significant. These effects are plotted in Fig. 41.

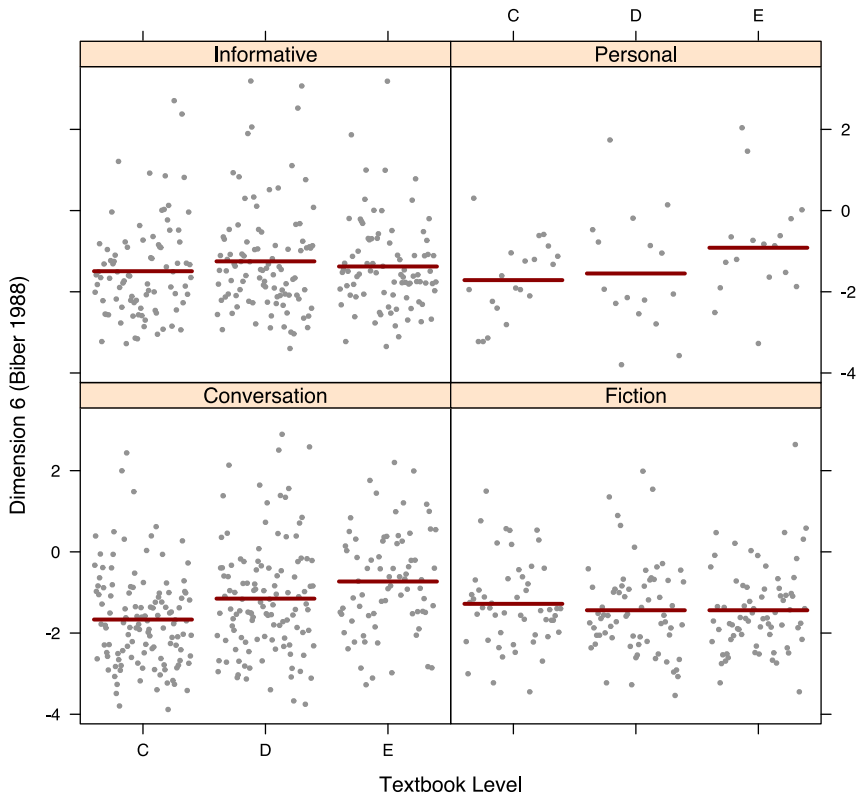


Fig. 41: Estimated scores on Biber's (1988) Dimension 6 across levels C to E textbooks

As seen in Fig. 41, the most noticeable textbook level effect arises in the Conversation register with a very moderate increase in Dimension 6 mean scores as learners are expected to become more proficient users of English. Extract (199) corresponds to an audio transcript from a level E textbook with a high Dimension 6 score (2.24). This is due to a relatively high density of complex sentence structures with *that*-clauses complementing verbs and adjectives, relative clauses, as well as its frequent use of demonstrative pronouns.

(199) Ben: [...] Australia Day should be about celebrating the present and the future, not worrying about things **that** happened 250 years ago.
Newsreader: Alice, can I bring you in at **this** point?
Alice: Obviously I disagree. Australia Day celebrates the arrival of the British on January the 26th 1788. A lot of people feel **this** is not something to be celebrated. Look at any group of protesters - it's not just black faces in the crowd. A very wide range of Australian people feel **that** the British colonisation of our land is not something that anybody should be particularly proud of. Many of us prefer to call it 'Invasion Day'.
Ben: Yeah, but don't you think it's right **that** people want to celebrate their Australian identity? Millions of people enjoy Australia Day. You really want to stop **that**?
Alice: **That**'s not what we're saying, and I think you know **that** really, Ben. <TEC: Green Line New 5>

6.3.2 Comparative additive MDA

Having examined the extent of register variation *within* school EFL textbooks, in the following, three major textbook registers, Conversation, Fiction and Informative writing, are compared with the three corresponding target language reference corpora on Biber's (1988) first three dimensions. In other words, the three reference corpora detailed in 3.3.2 are superimposed onto the additive MDA of the Textbook English Corpus (TEC) conducted in the previous section (6.3.1).

6.3.2.1 The specificities of Textbook English on Biber's (1988) Dimension 1

The distribution of the three TEC registers on Biber's (1988) Dimension 1 as compared to that of the three reference corpora is visualised in Fig. 42.

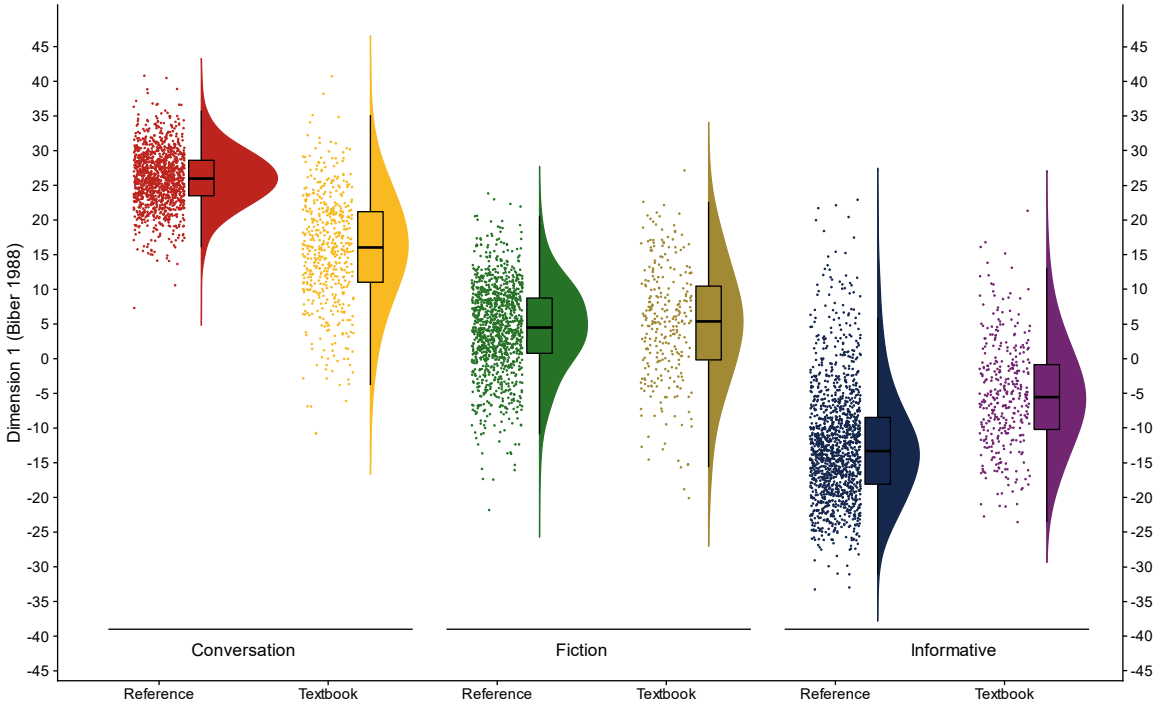


Fig. 42: Comparison of the Conversation, Fiction and Informative texts from the TEC with the three corresponding target language reference corpora on Biber's (1988) Dimension 1 (as calculated by the MAT, see below)

Although Textbook Conversation scores highest among the textbook registers, the Spoken BNC2014 displays considerably higher scores than Textbook Conversation ($\bar{x} = 15.75$, SD $= 7.89$ vs. $\bar{x} = 26.02$, SD $= 4.04$). Crucially, this difference is, in fact, even greater because most Spoken BNC2014 plotted on Fig. 42 are artificially deflated: this is caused by the absence of punctuation marks in the Spoken BNC2014. Indeed, the Biber Tagger and, as its faithful "copy", also the MAT requires the presence of punctuation marks to identify five of the 22 features with positive loadings on Biber's Dimension 1: stranded prepositions, discourse particles, non-phrasal clause coordination, sentence relatives and direct WH-questions. As a result, the MAT

returns counts of zero for sentence relatives and direct WH-questions in the Spoken BNC2014 because these features rely on commas and full stops, both of which are entirely absent from the corpus (Love, Hawtin & Hardie 2018: 37–38). Similarly, the discourse particle variable has an incredibly low mean normalised count of 0.05 (as compared to 0.36 in Textbook Conversation). Following Biber's (1988) operationalisation of this variable, only discourse markers preceded by a punctuation mark are counted as such by the MAT; thus, this low value only accounts for discourse markers occurring immediately after a question mark, as this is the only punctuation sign to occur in the Spoken BNC2014 (see example of the MAT output in (200)). Incidentally, this highly restrictive operationalisation of discourse markers also excludes many relevant features in Textbook Conversation, e.g., in (201) *well* is erroneously tagged as a noun. The same issue occurs with stranded prepositions, which are operationalised as prepositions immediately followed by a punctuation mark. Finally, the counts for non-phrasal clause coordination are also considerably deflated as its operationalisation also partly depends on punctuation.

(200) oh_UH is_VPRT [BEMA] he_TPP3 rapping_VBG ?_.
well_DPAR what_WP 's_VPRT [CONT] [BEMA] he_TPP3 doing_VBG [PROD] ?_.
<BNC2014: SQWN, as tagged by the MAT>

(201) Presenter_NN Oh_UH ?_. Why_RB 's_VPRT [CONT] that_DEMP ?_.
Gloria_NN **Well_NN** ,_, I_FPP1 'm_VPRT [CONT] [BEMA] from_PIN South_NN America_NN ,_, but_CC New_NN York_NN is_VPRT [BEMA] my_FPP1 new_JJ home_NN ._.
<TEC: Solutions Elementary, as tagged by the MAT>

Consequently, the five aforementioned features that rely on punctuation were excluded from the Dimension 1 scores of the Spoken BNC2014 and Textbook Conversation. New adjusted scores were calculated by adding the *z*-scores of the remaining features with positive loadings and subtracting those with negative loadings. Thus, the model summarised in Table 71 takes these adjusted comparable Dimension 1 scores as the outcome variable for the Textbook Conversation and the Spoken BNC2014 corpora. In Table 71 and the following model summaries, the reference levels are Textbook English for the Corpus variable and Conversation for the Register variable. In other words, the estimated Dimension 1 mean score for the Spoken BNC2014 (the reference corpus for Conversation) is 30.66, which corresponds to 16.01 plus the mean estimated for Textbook Conversation (14.65, the intercept). The estimated mean for the Info Teens is -12.20 which is the sum of the following coefficient estimates: 14.65 (intercept), 16.01 (Reference corpus), -20.72 (Informative register) and -22.15 (Reference*Informative interaction).

Table 56: Summary of the model: lmer(Dim1$_{adjusted}$ ~ 1 + Corpus + Register + Corpus*Register + (Register|Source))

Predictors	*Estimates*	*CI*	*p*-value
(Intercept) [Textbook Conversation]	14.65	12.29 – 17.02	**<0.001**
Corpus [Reference]	16.01	8.72 – 23.30	**<0.001**
Register [Fiction]	-10.75	-12.56 – -8.93	**<0.001**
Register [Informative]	-20.72	-22.70 – -18.74	**<0.001**
*Corpus [Reference] * Register [Fiction]*	-15.48	-22.64 – -8.33	**<0.001**
*Corpus [Reference] * Register [Informative]*	-22.15	-29.71 – -14.59	**<0.001**
	Random Effects		
σ^2	35.32		
$\tau_{00\ Source}$	12.34		
$\tau_{11\ Source.RegisterFiction}$	4.87		
$\tau_{11\ Source.RegisterInformative}$	7.32		
ρ_{01}	0.40		
	0.12		
ICC	0.35		
N_{Source}	325		
Observations	5033		
Marginal R^2 / Conditional R^2	0.829 / 0.889		

As shown on Fig. 43, the exclusion of the features that rely on punctuation for their operationalisation further widens the gap between naturally occurring conversation and textbook dialogues ($\bar{x}$ = 30.70, SD = 4.61 vs. $\bar{x}$ = 14.80, SD = 8.09, $\bar{x}$ estimated difference = -16.01, SE = 3.72, $p < .0001$).

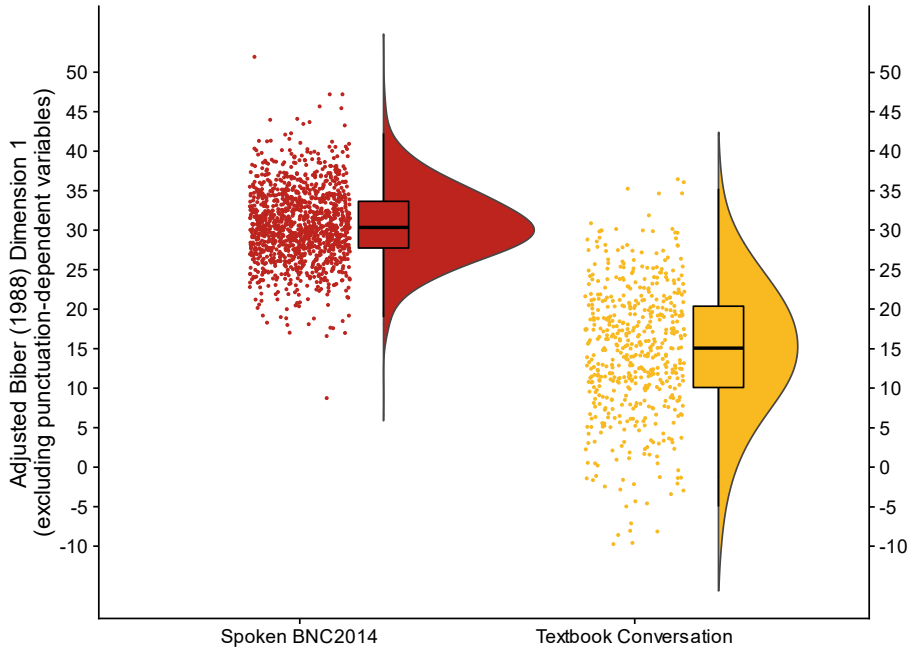


Fig. 43: Comparison of scores on Biber's (1988) Dimension 1 of the Spoken BNC2014 and Textbook Conversation, recalculated so as to exclude the features that rely on punctuation marks for their operationalisation

Table 57 sheds light on the linguistic features which most contribute to these characteristically low Dimension 1 scores for Textbook Conversation. Significance testing was performed with independent two-tailed Wilcoxon tests (asterisks denote p-values of $< .001$ after Holm correction). All the features listed in Table 57 except amplifiers, possibility modals, second-person pronouns and indefinite pronouns, contribute to textbook dialogues obtaining lower scores on this dimension.

Table 57: Normalised counts for the features loading on Biber's (1988) Dimension 1, except those that rely on punctuation for their operationalisations (features with positive loadings in red and bold, with negative loadings in blue)

Features that load on Biber's (1988) Dimension 1	Textbook Conversation		Spoken BNC2014		Comparison
	mean	*SD*	mean	*SD*	mean % difference
Hedges	0.11	*0.15*	0.26	*0.18*	-0.81*
***That*-deletions**	0.43	*0.35*	0.91	*0.26*	-0.72*
WH-clauses	0.12	*0.14*	0.22	*0.09*	-0.59*
Pronoun it	1.76	*0.80*	3.21	*0.66*	-0.58*
Nouns	24.40	*4.34*	14.44	*1.73*	0.51*
Causative subordination	0.14	*0.17*	0.23	*0.16*	-0.49*
***DO* as a pro-verb**	0.34	*0.29*	0.55	*0.21*	-0.47*
Emphatics	1.14	*0.61*	1.71	*0.51*	-0.40*
Analytic negation	1.55	*0.75*	2.19	*0.48*	-0.34*
Contractions	4.24	*1.53*	5.79	*0.86*	-0.31*
Amplifiers	0.31	*0.30*	0.23	*0.15*	0.30*
Demonstrative pronouns	0.76	*0.43*	1.02	*0.30*	-0.29*
Private verbs	2.00	*0.86*	2.56	*0.59*	-0.25*
Prepositions	6.58	*1.62*	5.40	*0.74*	0.20*
Type/token ratio	0.46	*0.05*	0.40	*0.03*	0.14*
Attributive adjectives	4.02	*1.33*	3.60	*0.60*	0.11*
Average word length	3.94	*0.26*	3.65	*0.10*	0.08*
Present tense	8.66	*2.27*	9.41	*1.22*	-0.08*
First-person pronouns	5.98	*2.14*	5.56	*1.19*	0.07*
Possibility modals	0.90	*0.53*	0.85	*0.28*	0.06
BE as a main verb	3.28	*1.07*	3.31	*0.45*	-0.01
Second-person pronouns	3.09	*1.41*	3.10	*0.84*	<0.00
Indefinite pronouns	0.05	*0.10*	0.05	*0.04*	<0.00

As compared to the Spoken BNC2014, the greatest underuses in Textbook Conversation are observed in the frequency of hedges (e.g., *sort of*), *that*-deletion (marked [THATD] in the example (202)) and the use of the pronoun *it*. Furthermore, WH-clauses (e.g., *do you know what I mean*), causatives (e.g., *because, cos*), DO as a pro-verb, emphatics (e.g., *just, really*), analytic negation, contractions, demonstrative pronouns and private verbs (e.g., *THINK, KNOW, BELIEVE, SEE, MEAN*) are also

considerably more frequent in naturally occurring conversation (202) than in textbook representations thereof (203).

(202) **it's** the the erm whatever you call it
greenfly
yes **it's** er s that **sort of**
greenflies
yes **it's it's** erm something from the greenflies I **think** rather than **it's not** the tree itself **it's**
the fact that **it's** the aphids erm producing something
do you **think** they drink too drink too much of **this** and **it** makes them ill?
I **think [THATD]** they go they go too too mad on the on the sap and **it just** produces all **this** sticky goo
oh gosh I **didn't know** <BNC2014: SRWD>

Nouns, on the other hand, appear to be considerably overrepresented in pedagogical dialogues (203). These high noun counts correlate positively with high frequencies of prepositional phrases, attributive adjectives, higher type/token ratios and longer words – all of which weigh negatively on this dimension. These features, together with relatively low frequencies of the features with positive loadings discussed above, frequently make textbook dialogues sound like rather unlikely transcripts of real-life conversations, e.g., (203).

(203) Man: Is that your **favourite British dish**?
Woman: Well, I like **roast beef** a lot. But my **real favourite** is waking up **in** the **morning** to the **smell** of a **full English breakfast**. Or **Welsh breakfast,** or the **full Irish breakfast**. Or the **Ulster fry**. Or **the Scottish breakfast**. **Eggs, bacon** and lots of other **tasty things**. It's more or less the **same** wherever you go in the **British Isles**. It's just the **name** that changes.
Man: Is that what you have for **breakfast** every **day**?
Woman: Well, not every **day**, but sometime **at weekends**. And of course, at **hotels** you can usually have the **full cooked breakfast** if you like. Tastes great with a **nice cup** of **tea**. By the way, did you know that **people** in the **British Isles** drink around three **kilos** of **tea** every **year**.
Man: Three **kilos**?
Woman: Yes, that's over ten **times** as much **tea** as **people** in **Germany** drink. Can you pass the **milk** and **sugar**, please? <TEC: Access G 3>

By contrast, textbook conversations with comparatively high Dimension 1 scores feature more verbal features, such as present tense forms, contractions, negation, first and second-person and *it* pronouns, as well as higher normalised counts of discourse markers, amplifiers, hedges, direct WH-questions and stranded prepositions than the majority of textbook dialogues (204).

(204) Jack: Lily, there**'s no way I'm** going to **recognise** a model, **it doesn't** matter how famous she **is**. But **I** tell **you what - I bet it isn't** her. What**'s** a famous model going to **be doing** in a shopping mall in **our** town?
Lily: **I think it is** her, you know! And she**'s going** into that shop. Come on - **let's** go in too.

Jack: **No way**. Even if **it is** her - leave her alone, she **just wants** to do some shopping. And **anyway**, what **are you going** to do - ask her for her autograph or something?
Lily: **I don't know**. **Maybe I'll just** go up and say hello. **What** do **you** reckon?
<TEC: English in Mind 4>

The model summarised in Table 71 does not lend support to the hypothesis that the dialogues featured in more advanced textbooks have higher, hence more authentic-like, Dimension 1 scores. In fact, some of the Level A textbook dialogues score comparatively high on Dimension 1 owing to their restricted vocabulary, shorter utterances and frequent turns leading to lower type/token and higher verb/noun ratios (205). By contrast, many of the texts intended to represent spoken interactions in the more advanced textbooks of the TEC are characterised by a much more nominal style with high informational density, thus featuring high type/token ratios, many prepositions and longer words (206).

(205) Lucy: **Hey**, watch **out**!
Sam: **Oh**, sorry! **Hey, you're** at Plymstock School.
Lucy: **So**?
Sam: **I'm** at Plymstock school too.
Lucy: **You aren't** from Plymouth!
Sam: **No**, **I'm not**. **I'm** new here. **I'm** from London.
Lucy: **OK**.
Sam: **I'm** in Year 7 in class 7EB. What about **you**?
Lucy: **I'm** in 7EB too.
Sam: **Hey**, that**'s** cool. <TEC: Access G 1>

(206) P: Thanks **for** your **input**, and **good luck**! Now, let's ask someone else. Hello, can I ask you what you think **of** the **American Dream**?
B: Hello! Well, my **ancestors** moved **to** the **United States** long ago, **in** 1846, **during** the **Irish potato famine**. They were **in dire straits** and wanted to escape **poverty**. They had to take care **of** themselves. They worked hard, and slowly they got richer and managed to build **a new life**. They saw the **US** as a **land of freedom** and **opportunity**, where everyone could work hard and be successful. <TEC: Piece of Cake 3e>

Textbook series only accounts for 8.4% of the large dispersion in Dimension 1 scores observed among Textbook Conversation texts. Nonetheless, the textbook dialogues of *English in Mind* (EIM) and *New Green Line* (NGL) tend to score marginally higher than the average textbook dialogue of the TEC. By contrast, the spoken representations of two of the three French textbook series, *Piece of Cake* and *Hi There*, score lowest on this modified version of Biber's (1988) Dimension 1. This is illustrated in Fig. 44, which shows how the effect ranges for estimated Textbook Conversation scores on this modified dimension compare (simulated using the R function merTools::REsim; Knowles & Frederick 2020).

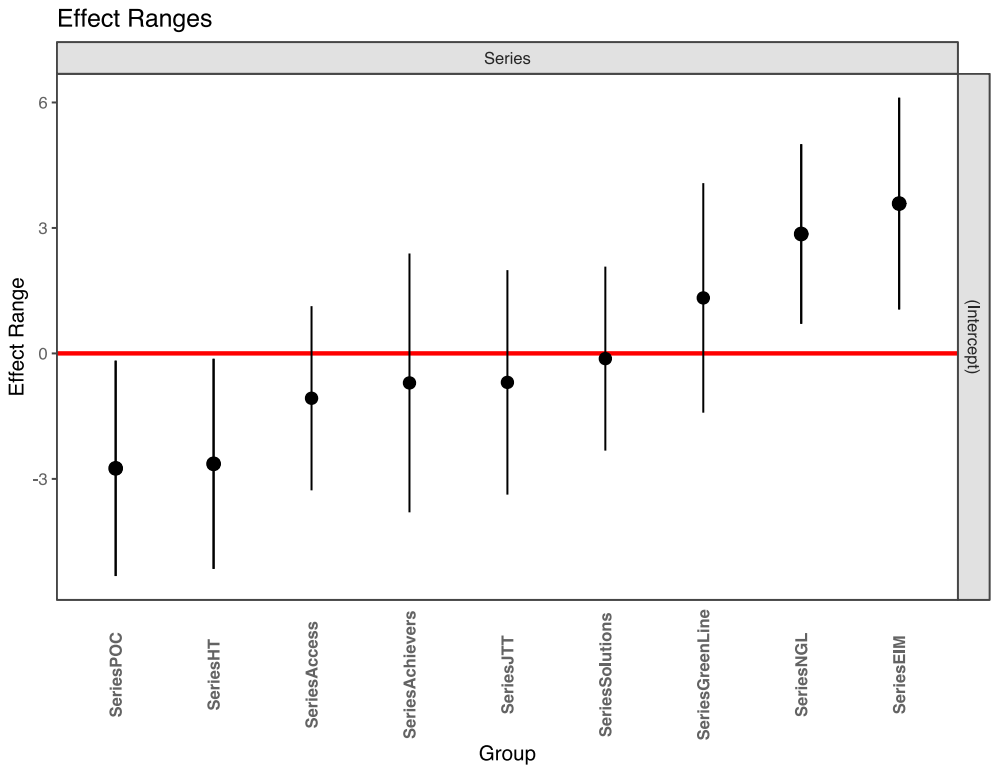


Fig. 44: Estimated varying effects of textbook series on modified Dimension 1 scores for Textbook Conversation (simulated using the R function merTools::REsim; Knowles & Frederick 2020)

Contrary to Textbook Conversation, which appears to be considerably less "oral-like" than the Spoken BNC2014 data, Informative texts in EFL textbooks tend to score higher on Biber's (1988) first dimension than the Info Teens corpus ($\bar{x}$ estimated difference $= 6.14$, $p = 0.002$). The features which most contribute to this mean difference are first and second-person pronouns, *DO* as a pro-verb, contractions, and amplifiers. The prevalence of these features reflects the often informal, "chatty" tone of the informative texts featured in school EFL textbooks (207).

(207) So how can **you** help yourself to remember things better in the long term? Well, there are several things **you** can **do**. One of them is to make sure **you** pay attention and take in the information properly in the first place. Others are to **do** with the effort you make to remember it afterwards. […] **Don't** wait to revise until exam time - by then **it's** too late!
Although the human brain is **amazingly** powerful, most people only use a tiny amount of its power. The brain is like a muscle. If **you don't** exercise it, it loses its strength and deteriorates. If **you** want to develop and improve **your** mind and make the most of it, **you** need to do regular mental exercises. In spite of all our potential brain power, **we** can easily forget 80% of what **we** learn in hours unless **we** make a special attempt to remember it. <TEC: Achievers B2>

The text from which (207) was extracted corresponds to the mean Dimension 1 score of the Textbook Informative subcorpus. By way of comparison, extract (208) scores

around the mean score of the Info Teens corpus. The latter is characterised by more nouns, prepositions, attributive adjectives and longer words.

(208) **Ayanna Pressley** has won her **election**, making her the **first black woman** to represent **Massachusetts** in the **House** of **Representatives**, **Boston.com** reports. She ran unopposed in **Massachusetts**'s 7th **district**.
Before the **polls** closed on **election day**, she urged **people** on **Twitter** to vote. "Today, we are powerful. There are only a **few hours** left to get out the **vote**. Go #vote for **progressive candidates** who will fight for **equity** and **justice**," she tweeted. "Vote for **activist leaders** who will work in and with **community**. Vote, because this is your **democracy** and your **voice** matters." <Info Teens: teenvogue.com>

In both the TEC and Info Teens, informative texts that score lowest on Dimension 1 tend to include bullet point lists and thus feature a high proportion of nominal sentences, as well as many attributive adjectives, a high type/token ratio and longer words, e.g., (209) and Fig. 45.

(209) **Borders**: **Ireland** is an **island** in the **North Atlantic Ocean**. The **Republic** of **Ireland** takes up most of the **island** with **Northern Ireland** (which is part of the **United Kingdom**) taking up a **northern section**.
Total Size: 70,280 **square km**
Size Comparison: slightly larger than **West Virginia**
General Terrain: mostly **level** to **rolling interior plain** surrounded by **rugged hills** and **low mountains**; **sea cliffs** on **west coast**
Geographical Low Point: Atlantic Ocean 0 **m**
Geographical High Point: Carrauntoohil 1,041 **m**
Climate: temperate maritime; modified by **North Atlantic Current**; **mild winters, cool summers;** consistently humid; overcast about half the **time**
<Info Teens: ducksters.com>

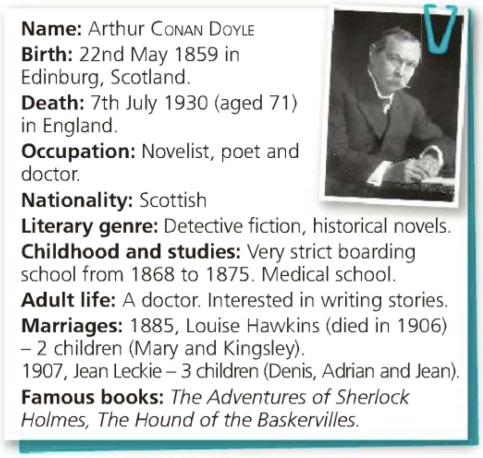

Name: Arthur Conan Doyle
Birth: 22nd May 1859 in Edinburg, Scotland.
Death: 7th July 1930 (aged 71) in England.
Occupation: Novelist, poet and doctor.
Nationality: Scottish
Literary genre: Detective fiction, historical novels.
Childhood and studies: Very strict boarding school from 1868 to 1875. Medical school.
Adult life: A doctor. Interested in writing stories.
Marriages: 1885, Louise Hawkins (died in 1906) – 2 children (Mary and Kingsley).
1907, Jean Leckie – 3 children (Denis, Adrian and Jean).
Famous books: *The Adventures of Sherlock Holmes, The Hound of the Baskervilles.*

Fig. 45: Informative textbook text with a low Dimension 1 score <TEC: Join the Team 4e>

Contrary to the two textbook registers discussed above, the difference in mean Dimension 1 scores between Textbook Fiction and the reference Youth Fiction is not significant ($\bar{x}$ estimated difference -0.53, SE = 1.71, $p = 0.78$). Fiction usually consists of alternating narration and fictional speech. Thus, novels with a high

proportion of dialogues inevitably score high on Biber's first dimension, whilst those with longer descriptive passages score lower. Indeed, additive MDAs of 19th century novels have shown large significant differences on Biber's Dimension 1 between narrative passages, which are more associated with features corresponding to the informational end of the scale, and fictional speech, which is more associated with features characteristic of involvement and interaction ($\eta^2 = .83$, $p<0.001$; Egbert & Mahlberg 2020: 85; cf. Biber & Finegan 1994). These findings imply that this dimension is not best suited to examine the potentially defining characteristics of Textbook Fiction. That said, the non-significant difference in mean Dimension 1 scores for Textbook Fiction and the Youth Fiction does suggest that they feature similar proportions of narration to fictional speech.

In addition, the model estimates for the mean Dimension 1 scores of the TEC registers listed in Table 71 make clear that the small but significant effect of textbook level on Dimension 1 scores is driven by its interactions with the Fiction register: Textbook Fiction tends towards marginally lower Dimension 1 scores as the proficiency level of the textbooks increases. As mentioned in 6.3.1.1, this finding must be approached with caution: not only are the effect sizes very small, Fig. 32 also showed that some series feature no or few Fiction texts at certain levels. Nonetheless, we can observe that beginner textbooks tend to feature more dialogue-heavy fictional writing, leading to a greater use of first and second-personal pronouns, contractions, negation and demonstrative pronouns (210), than more advanced teaching materials, which often feature many more prepositions, nouns and attributive adjectives (211). Moreover, beginner textbooks that have yet to introduce past tense forms rely on present-tense narration, which also contributes to these higher Dimension 1 scores (210).

(210) 'Very funny,' Lucy **says**. '**I** think **this** is just a silly trick. **I don't** believe a word.'
'A silly trick?' the Time Lord **laughs**. 'Ha, ha, ha, just look at **this**, **you** silly girl!'
The lights in the Planetarium **flicker** again, and on the huge screen, Lucy, Sandy and Asim can see pictures of Greenwich - and it already **looks** very different. There aren't many old people any more, and children **are looking** down at clothes that **are** too big for them.
Then they **hear** the scary voice again.
'So, children. The future of the human race lies in **your** hands. See **this** hourglass here? When the sand is through, **your** time will be up. [...]'
<TEC: Green Line New 1>

(211) The **mountains** stretched **into** infinity: **exquisite shades** of green, grey and brown **against** a **deep azure, cloudless sky**. **Along** the **wall**, here and there, were **small groups of tourists basking in** the **wonder of** their **surroundings**. But the **strangest sight of** all was a **table** and four **plastic chairs** beneath a **huge red parasol**, and a man selling **bottled water** and **cans of chilled drinks from** an **icebox**. <TEC: Achievers B2>

6.3.2.2 The specificities of Textbook English on Biber's (1988) Dimension 2

Biber's second 'Narrative vs. Non-narrative Concerns' dimension is also of particular interest to the analysis of Textbook English as many textbook texts can be said to have a narrative function. Fig. 46 compares the Conversation, Fiction and Informative texts of the TEC with the three reference corpora: it immediately makes apparent that the texts of Info Teens display the greatest variation in their degree of narrativeness. In fact, although their mean scores are very similar ($\bar{x}$ = -1.76 vs. -1.95), the Info Teens and the Spoken BNC2014 evidently follow very different distributions on this dimension. Textbook Fiction scores highest among the textbook registers but, on average, the texts of the Youth Fiction corpus score considerably higher ($\bar{x}$ = 2.53, SD = 2.83 vs. $\bar{x}$ = 4.91, SD = 2.08).

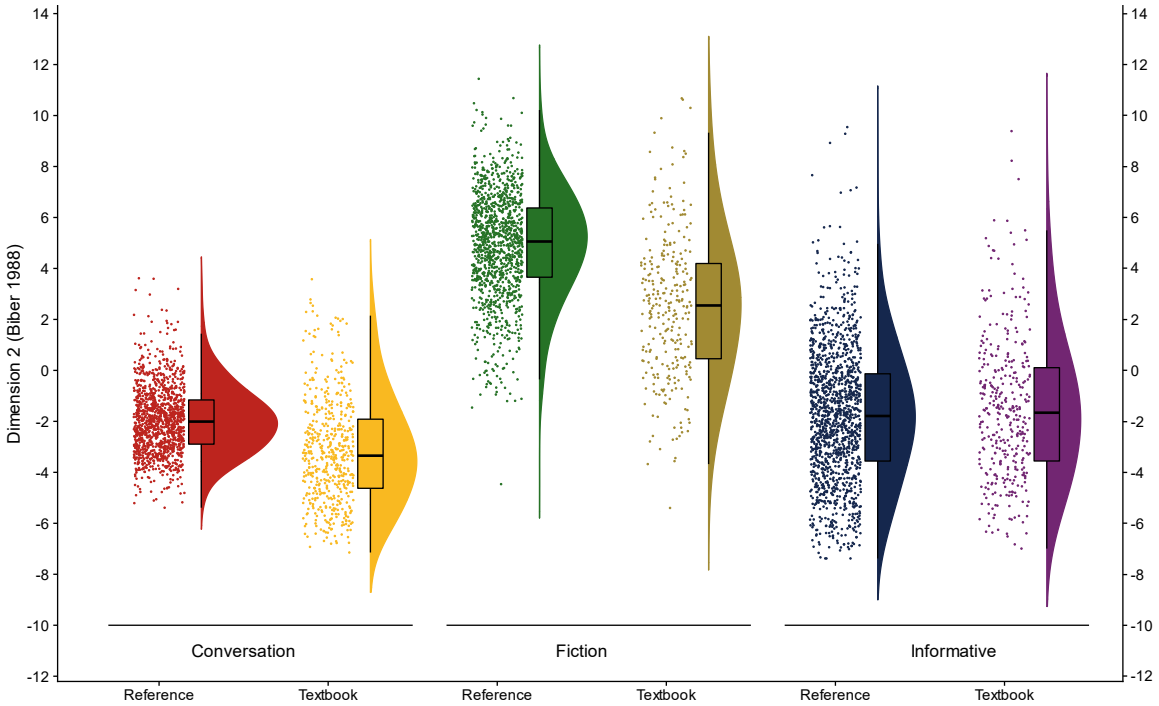


Fig. 46: Comparison of the Conversation, Fiction and Informative texts from the TEC with the three corresponding target language reference corpora on Biber's (1988) Dimension 2

Table 58 summarises the model computed to predict scores on Biber's (1988) Dimension 2 for the three TEC subcorpora and the three corresponding reference corpora. As in Table 71, the reference levels are Corpus: Textbook and Register: Conversation.

Table 58: Summary of the model: lmer(Dim2 ~ 1 + Corpus + Register + Corpus*Register + (Register|Source))

Predictors	*Estimates*	*CI*	*p*-value
(Intercept) [Textbook Conversation]	-3.10	-3.47 – -2.74	**<0.001**
Corpus [Reference]	1.15	0.14 – 2.17	**0.026**
Register [Fiction]	5.75	4.51 – 6.98	**<0.001**
Register [Informative]	1.37	0.68 – 2.06	**<0.001**
*Corpus [Reference] * Register [Fiction]*	1.10	-0.47 – 2.68	0.169
*Corpus [Reference] * Register [Informative]*	-1.15	-2.45 – 0.16	0.084
	Random Effects		
σ^2	3.67		
$\tau_{00\ Source}$	0.23		
$\tau_{11\ Source.RegisterFiction}$	3.23		
$\tau_{11\ Source.RegisterInformative}$	0.92		
ρ_{01}	-0.58		
	0.1		
ICC	0.25		
N_{Source}	325		
Observations	5033		
Marginal R^2 / Conditional R^2	0.649 / 0.738		

Table 46 lists the estimated means for each the three TEC and reference registers. Of these three comparisons, the Conversation ($p = .03$) and Fiction ($p < .001$) contrasts are significant. The Textbook Informative and ITCC texts, on the other hand, have remarkably similar estimated mean Dimension 2 scores.

Table 59: Estimated Dimension 2 means (degrees-of-freedom method: asymptotic)

	mean	**SE**	**lower CL**	**upper CL**
Textbook Conversation	-3.10	0.19	-3.47	-2.74
Spoken BNC2014	-1.95	0.49	-2.90	-1.00
Textbook Fiction	2.65	0.55	1.56	3.73
Youth Fiction	4.90	0.11	4.69	5.11
Textbook Informative	-1.73	0.39	-2.50	-0.97
Info Teens	-1.73	0.30	-2.31	-1.16

The intra-textbook analysis in 6.3.1 demonstrated that this dimension is subject to strong textbook level effects because it comprises grammatical features, such as the past tense and perfect aspect, which are usually not taught until the second or third year of English instruction at secondary school. A comparison of the present results with

Table 50, which summarises these textbook level effects, shows that, on average, the dialogues of the more advanced textbooks score very similarly to the Spoken BNC2014 on this narrative dimension.

However, although Dimension 2 means for Textbook Fiction do dramatically increase from -0.15 to 2.92 between Level A and Level B textbooks to reach a peak in Level C textbooks (3.48), even that latter score remains significantly lower than the Youth Fiction estimated mean (4.90). Fig. 47 explores the reason behind this lower average score for Textbook Fiction as compared to the Youth Fiction reference data by comparing the relative frequencies of all six features that load on Dimension 2. As already explained in 6.3.1.2, four of these features are largely absent from beginner textbooks; however, the boxplots for past tense, perfect aspect verbs, and present participial clauses (see Fig. 47) suggest that the median frequencies at the more advanced levels are close to the Youth Fiction ones. It is also worth noting that for particularly rare features such as synthetic negation and present participial clauses, the difference in means may well be largely attributable to text length only: indeed, textbook fictional texts are, on average, considerably shorter than the text samples of the Youth Fiction and this reason alone may account for the apparent absence of these features in Textbook Fiction texts. By contrast, the trend for public verbs across textbook levels counters expectations: whilst Level A and B fictional texts feature above-average relative frequencies of public verbs, foremost *SAY*, *WRITE, EXPLAIN, REPLY, REPORT* and *EXPLAIN*, Level D and E texts tend to feature fewer than in the Youth Fiction reference corpus. This confirms the finding that beginner Textbook Fiction texts feature more dialogues than narrative passages.

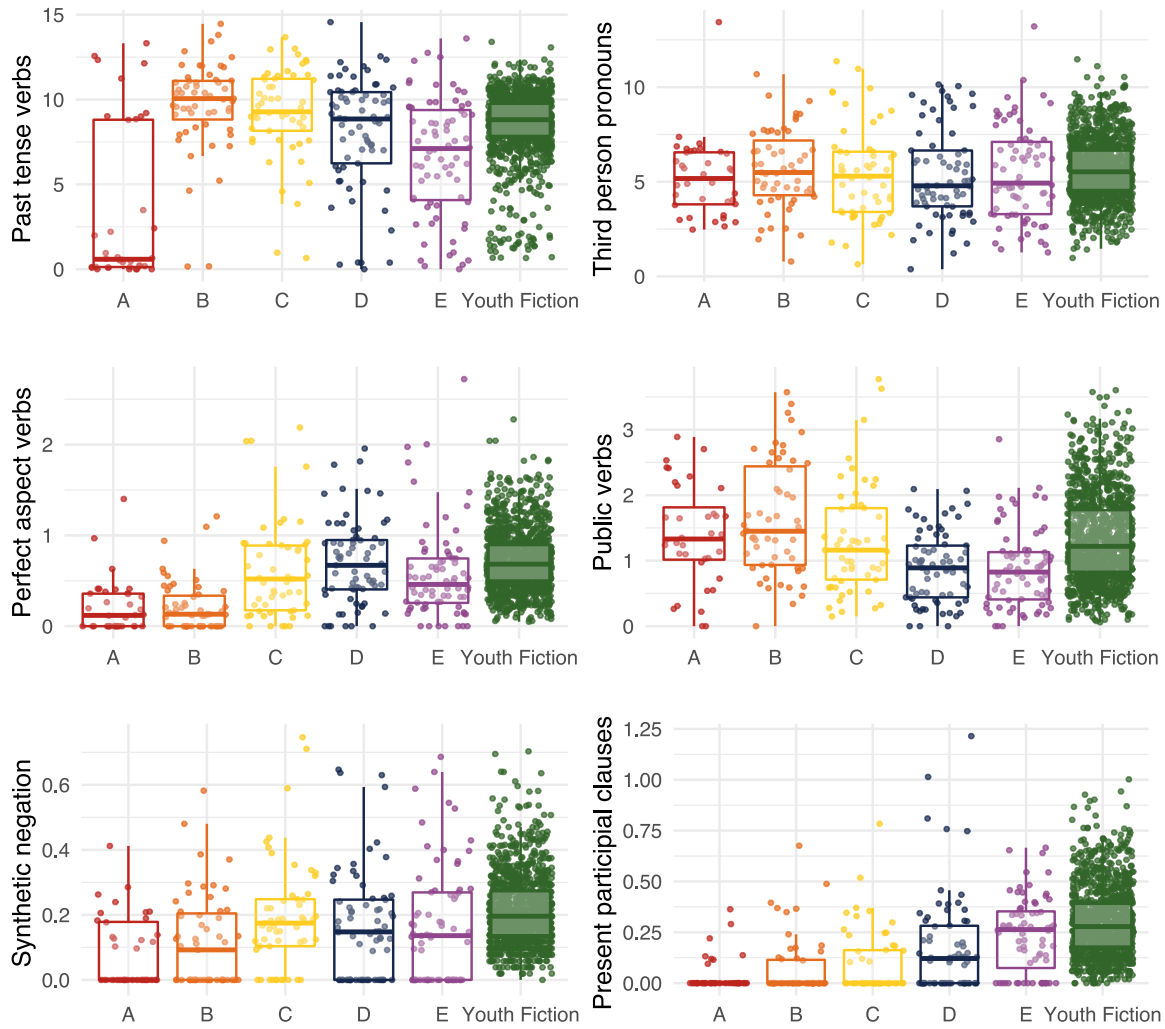


Fig. 47: Normalised frequencies of features loading on Biber's (1988) Dimension 2 for Textbook Fiction (subdivided by proficiency level A to E) and Youth Fiction

6.3.2.3 The specificities of Textbook English on Biber's (1988) Dimension 3

On Biber's (1988) third dimension, the Spoken BNC2014 and Youth Fiction texts largely score negative values and follow very tight distributions ($\bar{x}$ = -1.18, SD = 1.29 and $\bar{x}$ = -0.985, SD = 1.62), whereas the mean score for the Info Teens is positive and the texts show much greater variation ($\bar{x}$ = 4.18, SD = 3.82).

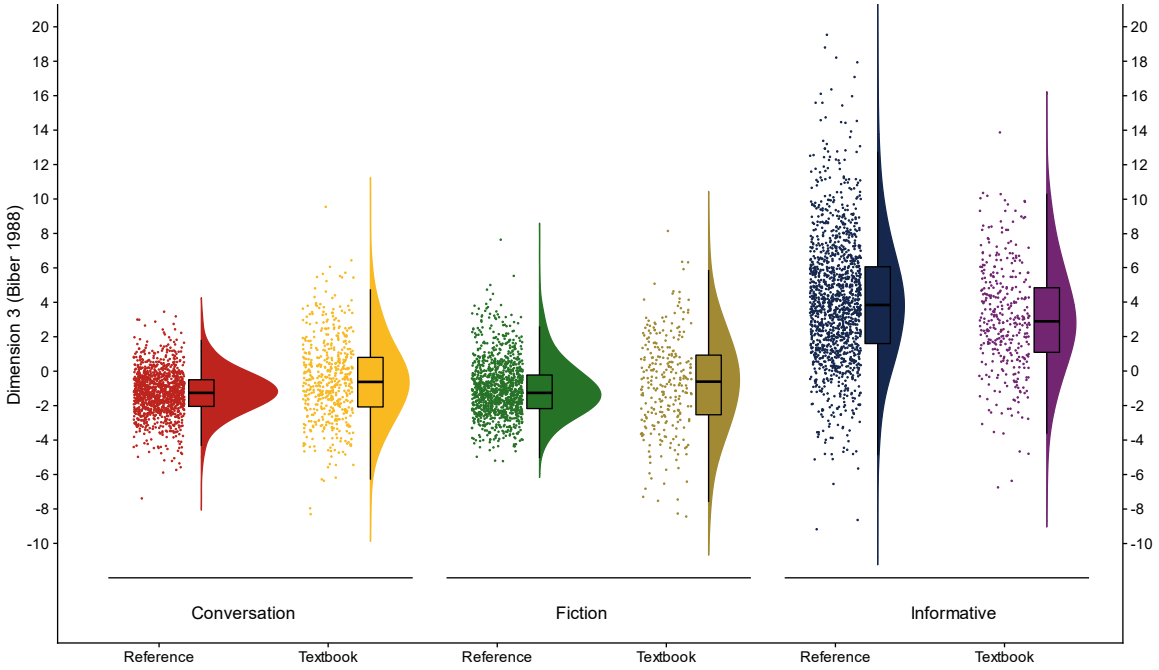


Fig. 48: Comparison of the Conversation, Fiction and Informative texts from the TEC with the three corresponding target language reference corpora on Biber's (1988) Dimension 3

As models with both random intercepts and slopes failed to converge, Dimension 3 scores were modelled with a simplified random effect structure with random by-source intercepts only (see 6.2.6). As shown in Table 60, only one significant effect can be observed on this dimension.

Table 60: Summary of the model: lmer(Dim3 ~ 1 + Corpus + Register + Corpus*Register + (1|Source))

Predictors	*Estimates*	*CI*	*p*-value
(Intercept) [Textbook Conversation]	-0.40	-1.09 – 0.28	0.248
Corpus [Reference]	-0.78	-2.85 – 1.29	0.462
Register [Fiction]	-0.06	-0.43 – 0.30	0.733
Register [Informative]	3.74	3.41 – 4.08	**<0.001**
*Corpus [Reference] * Register [Fiction]*	0.26	-1.73 – 2.25	0.799
*Corpus [Reference] * Register [Informative]*	1.61	-0.44 – 3.66	0.123
	Random Effects		
σ^2	5.65		
$\tau_{00\ Source}$	0.99		
ICC	0.15		
$N_{\ Source}$	325		
Observations	5033		
Marginal R^2 / Conditional R^2	0.459 / 0.539		

Table 61 lists the estimated Dimension 3 means allowing for a direct comparison between the Conversation, Fiction, and Informative texts of the TEC and the three

corresponding reference corpora. In addition, pairwise comparisons show that none of these reported mean differences between the three TEC and reference registers are statistically different. Thus, on this dimension, the three TEC registers can be said to largely match their target language counterparts. Note, however, that, as with the other dimensions, register differences appear to be less marked within textbooks than across the three target language reference corpora. Indeed, even though the differences are not significant, Table 61 suggests that, on this dimension, the conversations of the Spoken BNC2014 tend to score lower than Textbook Conversation, whereas the texts of the Info Teens usually score higher than Textbook Informative texts.

Table 61: Estimated mean Dimension 3 scores (degrees-of-freedom method: asymptotic)

	mean	**SE**	**lower CL**	**upper CL**
Textbook Conversation	-0.40	0.35	-1.09	0.28
Spoken BNC2014	-1.18	1.00	-3.13	0.77
Textbook Fiction	-0.47	0.36	-1.18	0.25
Youth Fiction	-0.99	0.09	-1.16	-0.81
Textbook Informative	3.34	0.36	2.64	4.04
Info Teens	4.17	0.27	3.65	4.70

6.4 Discussion

This chapter presented the results of two additive MDAs using Biber's (1988) model of general spoken and written English as the baseline model. It demonstrated that, with some caveats, additive MDA can be successfully used to explore linguistic variation within school EFL textbooks. In the first half of the results section (6.3.1), the major registers of the TEC were compared to each other along each of Biber's (1988) six dimensions.

In answer to this chapter's RQ1, these additive MDAs have shown that Textbook English does vary considerably across different text registers. This was most striking on Dimension 1, where register explains 63.0% of the variance observed in dimension scores across the five major registers of the TEC (Conversation, Fiction, Personal correspondence, Instructional and Informative texts). By contrast, on this same dimension, textbook level and textbook series only explain 2.4% and 5.3%, respectively. This is an important finding, which confirms the absolute necessity to consider register as a key factor when investigating and/or evaluating the language of EFL textbooks – an aspect which, as shown in the literature review (Chapter 2), has largely been ignored to date. However, the results made clear that not all of Biber's (1988) dimensions are equally useful for exploring linguistic variation in EFL textbooks. Whilst register also accounts for 40.6% of variance on Dimension 2, 33.7% on Dimension 3, and 21.7% on Dimension 5, it explains less than 2% along Dimensions 4 and 6. In addition, Dimensions 4, 5 and 6 comprise some relatively rare lexico-grammatical features, which are difficult to reliably account for in the often

relatively short texts of the TEC – an issue which persisted even after having concatenated texts totalling fewer than 400 words, as explained in 6.2.2.

Moreover, the six models computed as part of the intra-textbook analysis (see 6.3.1) confirmed that some of Biber's (1988) dimensions are more subject to textbook level effects than others. In particular, the Dimensions 1 and 3 scores of the TEC texts are relatively stable across the five different textbook levels, whereas some salient proficiency level effects were observed along Dimensions 2, 4, and 5, accounting for 6.5%, 14.7%, and 14.6% of variance among these dimension scores, respectively. In many instances, significant interactions between textbook register and textbook level were identified.

Overall, the mixed-effects models showed that Textbook English, as a language variety, is relatively homogenous across the nine different textbook series and three countries of use represented in the Textbook English corpus. The largest, though still very modest, differences between the textbook series and countries of use were observed on Dimensions 1 and 3, on which textbook series accounted for 5.3% and 13.9% of variance, respectively. In both instances, it was French textbook series that stood out: on Biber's (1988) Dimension 1, the dialogues of the series *Piece of Cake* and *Hi There* were found to score significantly lower than the Textbook Conversation average, and, on Biber's (1988) Dimension 3, all three French series scored higher across all textbook registers, thus suggesting that their texts are more explicit and less situation-dependent than those of their Spanish and German counterparts.

Whilst Section 6.3.1 showed considerable intra-textbook register variation (which largely followed the expected patterns along Biber's (1988) Dimensions 1, 2, 3 and 5), it would nevertheless appear that register-based linguistic differences are far less pronounced within textbooks than across the different text registers entered in Biber's (1988) model. This trend was confirmed in the comparative MDA analysis conducted in 6.3.2. which addressed RQ2 and RQ3: *To what extent do Textbook English registers differ from target language registers and what are the defining linguistic features that characterise Textbook English registers as compared to these target language registers?* To this end, the three TEC registers with, on average, the longest texts, Fiction, Informative and Conversation, were compared to three target language reference corpora (see 3.3.2). For instance, on Biber's (1988) Dimension 1, Textbook Conversation scored considerably lower than the Spoken BNC2014, whilst Textbook Informative texts, on average, scored higher than the texts of the Info Teens. Thus, register-based linguistic variation in Textbook English appears to be more constrained than in situationally similar, naturally occurring registers.

When accounting for textbook level effects, Textbook Fiction resembles most closely its corresponding reference corpus of Youth Fiction novels. Two reasons were

proposed as to why fictional texts from beginner textbooks score higher than Youth Fiction on Biber's (1988) Dimension 1. First, it may be the result of a higher proportion of direct speech to narration in texts targeted at beginner learners. However, it is also certainly due to the fact that, before the past tense is introduced to learners, fictional narration is necessarily conducted in the present tense. This second cause is also clearly observable on Biber's (1988) second, 'Narrative vs. Non-narrative Concerns', dimension, on which beginner Textbook Fiction scores lower than Youth Fiction. Encouragingly, however, as soon as the past tense and perfect aspect have been introduced, few of the originally observed differences between Textbook Fiction and the Youth Fiction remain significant.

The informative texts of the TEC, on the other hand, were shown to differ more substantially from their counterparts of the Info Teens. On Biber's (1988) Dimension 1, they were found to be more interactional and spoken-like than the texts featured on informative websites targeted at English-speaking teenagers, i.e., they feature considerably more present tense verbs, contractions, and first and second-personal pronouns. In addition, textbook authors, understandably and likely consciously, do not employ many of the lexico-grammatical features that characterise informative texts (e.g., the feature that load on Dimension 4: passives, past participial clauses, conjuncts, and adverbial subordinators) until learners are expected to have mastered simpler structures. As a result, on several dimensions, the informative texts of beginner textbooks are significantly less like the corresponding target language reference corpus than those printed in the more advanced textbooks.

The most striking TEC vs. reference corpora differences were observed in the Conversation register. On Biber's (1988) Dimension 1, in particular, Textbook Conversation scores markedly lower than the Spoken BNC2014. This is largely due to the much more nominal style of textbook dialogues, which tend to feature longer speaker turns, longer words and higher type/token ratios. Thus, textbook dialogues appear to primarily function as reinforcers of previously acquired vocabulary, rather than as models of realistic spontaneous spoken interactions. Excluding the features that rely on punctuation for their operationalisation, the most underrepresented Dimension 1 features in Textbook Conversation were found to be hedges, *that*-deletions, WH-clauses and *it* pronouns.

In spite of some limitations (see 6.5), this chapter has shown the potential of additive MDA to both explore linguistic variation within textbooks and compare individual textbook registers with comparable target language registers along major functional dimensions of variation. It also has the advantage of yielding results that are potentially comparable to many other additive MDAs based on Biber's (1988) model of general English. That said, the main advantage of the method presented in this chapter undoubtedly lies in the relative simplicity of conducting additive MDAs based

on Biber's (1988) model thanks to the availability of the MAT (Nini 2014) that largely automates the process (see 6.2.3). Indeed, the MAT also includes a graphical user interface (for Windows only), thus genuinely making the method accessible beyond academia. It is therefore hoped that the method may be of interest to textbook authors, editors, publishers, and representatives of educational authorities who may want to consider applying additive MDA using the MAT as part of a wide range of methods for textbook evaluation and revision purposes.

Though it is by no means claimed that additive MDA could or should be used as a unique solution, this chapter has shown that some of Biber's (1988) dimensions can provide a valuable synthesis of the relative frequencies of relevant linguistic features that can help to distinguish particularly unnatural-sounding textbook texts from more natural-sounding ones. In particular, we saw that Biber's (1988) Dimension 1 lends itself particularly well to the examination of representations of spoken language as it captures functional variation along an involved/oral vs. informational/literate continuum. Thus, a high score on Biber's Dimension 1, such as that scored by the dialogue quoted in excerpt (212) (Dim 1 = 38.19 as calculated by the MAT; items that contributed to this high score are highlighted in bold), points to a pedagogical text that is likely to paint a more authentic picture of natural conversation than one with a much lower score, e.g. (213).

(212) Amy: **Hi**, Nick.
Nick: **Hi**, Amy. Amy, **is** this **your** backpack on the floor?
Amy: That**'s** right.
Nick: **Well**, **could you perhaps** put **it** somewhere else? **It's kind of** in the way.
Amy: **No, it's not. It's** where **I** always **leave it**.
Nick: **Yes, I know you** always **leave it** there. And **it's always** in the way. This **is** a pretty small place, Amy. So **perhaps** just for once **you could** put **your** backpack somewhere where **it isn't** in the way, **hmm**?
Amy: **You don't own** this place, Nick. So **don't try** and **tell me** what to do. **I** came in early to get some things done. **I** put **my** backpack on the floor. **You deal** with **it**! <TEC: English in Mind 4>

(213) Journalist: This is **Sally Gordon** here **in Leicester Square, London**. I'm right in the **middle of sports fans**. Excuse me, **Sir**. Who is your **favourite sports hero**?
Dwayne: Definitely, **Chris Hoy**, the **British track cyclist** - won two **gold medals**. He represents **strength** and **courage**, he never gave up.
Journalist: What about you? Who is the **best representative of** your **country**?
Donna: **Kobe Bryant** for sure. I'm American and we are very patriotic when it comes **to sport**. He has shown the **world** we remain **the dominant leaders in basketball**, no doubt. And **Michael Phelps** of course.
Journalist: Why?
Donna: Why? He has just won four **golds** and **two silver medals** and he is a **record holder**. The **dream** came true. Incredible. That's why he is nicknamed "the **Baltimore Bullet**". He symbolises **determination**,

generosity, hope... great **values**. You see, he's a **role model**! He will be remembered forever. <TEC: New Mission 2e>

For example, excerpt (213) scored -6.10 on Biber's (1988) Dimension 1, which is the result of its considerably higher type/token ratio and longer average word length than most natural conversations, as well as the fact it features many complex nominal phrases, which lead to high relative frequencies of prepositions and attributive adjectives – all of which contribute negatively to Dimension 1 scores. Thus, textbook dialogues that score particularly low could be flagged as potentially worth re-examining or revising (see Le Foll 2021b). Inversely, when textbook informative texts score particularly high on Biber's (1988) Dimension 1, this is a sign that they are unlikely to be of use as models for students to acquire the skills necessary to write their own informative texts or read for information independently outside the classroom; hence, here too, corpus-informed revisions should be considered (see Chapter 8).

6.5 Limitations

The present chapter also highlighted several issues in applying Biber's (1988) model as the baseline in an additive MDA aimed at exploring linguistic variation in school EFL textbooks and comparable target language registers. Solutions to overcome issues related to text length (see 6.2.2), the non-independence of texts from the same textbook series, web domain or novel (see 6.2.6), and the punctuation-dependent operationalisation of some of the features entered in Biber's (1988) MDA (see 6.3.2.1) were discussed and implemented. Mixed-effects models were used to identify and quantify additional factors of linguistic variation, such as the proficiency levels, series, and countries of publication/use of the textbooks (see 6.2.6). In comparison with text register, these were shown to only play a marginal role in mediating textbook language variation.

For a more precise and robust analysis of Textbook English registers, however, a full MDA is conducted in Chapter 7. This new multi-feature/multi-dimensional model aims to overcome some of the limitations encountered in the application of additive MDA to the study of Textbook English. First, this chapter has shown that there is clearly a need to carefully consider which lexical, grammatical and/or semantic features are most relevant for the linguistic analysis of school EFL textbooks. Biber's (1988) Dimensions 5 and 6, in particular, were shown to be of very limited use because they capture features which are rare in general English and would thus require considerably longer texts to be measured with any degree of accuracy.

Crucially, this chapter has also shown that the operationalisation of a number of Biber's (1988) features is not well suited to the present data. Thus, in addition to choosing salient features, it must also be ensured that the selected features can be reliably identified using automatic means because, given the large number of texts

and features to be tagged and counted, even partial manual annotation is clearly not feasible. This chapter has pinpointed a number of issues with the feature operationalisations applied in Biber's (1988) MDA. The exclusive reliance on punctuation marks to identify linguistic features such as discourse markers and stranded prepositions, which are highly frequent in transcripts of spoken language that may not include any punctuation, is a case in point. In addition, the discussion of register variation along Biber's (1988) Dimension 3 (see 6.3.1.3) revealed that, applying the original operationalisation, the nominalisation variable captures typical and highly frequent classroom nouns, such as *activity* and *document,* which inevitably skew the results. The nature of ELT tasks may also call for a re-consideration of the semantic verb categories used (which Biber borrowed from Quirk et al. 1985: 1180–1883). Indeed, the private verb class includes *BELIEVE, FEEL* and *PRETEND,* but also *KNOW, LEARN* and *SHOW,* and the public verb class includes many reporting verbs such as *CLAIM, SAY, REPORT* and *WARN,* but also *EXPLAIN, PRONOUNCE, REPEAT* and *WRITE*, which are likely to fulfil rather different functions in textbook language. Similarly, there is a risk that the results concerning the comparison of Textbook Conversation vs. the Spoken BNC2014 representing naturally occurring conversation were skewed because Biber's (1988) noun variable aggregates common and proper nouns and many textbook dialogues include the name of the person speaking at the start of every turn, e.g., (204)–(205). This will undoubtably have inflated the relative frequencies of nouns in these textbook dialogues and an improved method will need to remedy this.

As a result of these aforementioned limitations that arose from following the traditional MDA framework and Biber's (1988) model as a baseline, the following chapter lays out an improved method with the aim of developing a more accurate and robust model of linguistic variation in Textbook English.

7 Towards a new multi-dimensional understanding of Textbook English

I hope to travel and learn some languages. Maybe I'll become a journalist. Or an interpreter. Or even an English teacher. That will be tough - I don't know enough about English grammar!
<TEC: Achievers B1+>

7.1 Introduction

The present chapter addresses the same three sets of research questions pursued in Chapter 6:

1. What is the extent of the linguistic variation across the major registers of Textbook English? How are the different textbook registers characterised linguistically? To what extent do the proficiency levels of textbooks interact with register-based variation? Do some textbook series show significantly more or less register-based variation?
2. To what extent do Textbook English registers differ from situationally similar, naturally occurring registers? To what extent are (some of) the observed patterns moderated by textbook series, their country of use, and/or the proficiency level of individual textbook volumes?
3. What are the defining linguistic features that characterise Textbook English registers as compared to these target language registers?

Chapter 6 already provided initial answers to these questions and concluded with a summary of the strengths and limitations observed in applying Biber's (1988) model of general spoken and written English to the study of Textbook English. Whilst Chapter 6 followed an 'additive MDA' approach (Berber Sardinha et al. 2019) to the study of Textbook English as a variety of English, the present chapter tackles the aforementioned research questions by conducting a 'full' or 'novel' MDA (see 6.1.2.2 for an explanation of the distinction) with the aim of providing more comprehensive, detailed and robust answers to these questions.

7.2 Methodology: the MDA framework revised

Whilst the MDA framework (Biber 1984; 1988; 1995) remains a highly influential and popular method of multi-feature linguistic variation analysis (see, e.g., Berber Sardinha & Biber 2014; Berber Sardinha & Veirano Pinto 2019; Goulart & Wood 2021), it is not without its problems. Some of the most comprehensive criticism has been expressed by Evert (e.g., 2018) and Lee (2000). Some of these issues have already been touched upon in 6.5.

Potential methodological pitfalls associated with the MDA paradigm (as it is most frequently applied) can be summarised as:

- Design bias in the selection of text samples
- Design bias in the choice and operationalisation of features
- Uncertainty over the reliability of the feature counts
- Confusion between covariation of features due to situational variation and covariation due to grammatical structure
- Lack of transparency in the quantitative patterns captured by factor analysis
- Degradation of correlations due to poorly distributed variables
- Arbitrary thresholds in the computation of dimension scores
- Misleading visualisation of the results
- Difficulties in establishing the statistical significance and robustness of the results
- Issues with the reproducibility and replicability of the results

The following section addresses these ten areas of potential issues, outlining how they have guided the various linguistic, computational, and statistical decisions made as part of the design of the multi-feature/multi-dimensional method applied in the present chapter. In sum, this methodology is strongly inspired by Biber's MDA framework (see 6.1.1) but departs from the way it is traditionally applied in a number of significant ways. Modifications to the framework have been implemented both as a result of general, methodological issues associated with MDA (as outlined by Evert 2018; Lee 2000 and others) and of specific problems arising from the nature of Textbook English and the research questions outlined in 7.1.

Section 6.1.1 explained how, in unsupervised approaches to situational variation analysis such as MDA, the characteristics of each text in a corpus are first summarised in a numeric vector, which stores the relative frequencies of a large number of linguistic features in any one text. With this in mind, two potential design biases are evident: the selection of text (samples) to be analysed and the choice and operationalisation of linguistic features to be counted in these texts. The first issue is discussed in 7.2.1, whilst the latter is covered in 7.2.2. This is followed by a section discussing issues related to the (often unquantified or, even, unquantifiable) reliability of these feature frequencies and how it has been dealt with in the present chapter (7.2.3). Issues arising from the blanket use of per-word normalisation baselines are addressed in 7.2.4. Sections 7.2.5 and 7.2.6 shed light on some of the potential pitfalls associated with different factor analysis methods and, in particular, exploratory factor analysis (EFA) with oblique rotation of the factor matrix and explain why principal component analysis (PCA) was chosen instead. The next two sections deal with potential pitfalls in the interpretation (7.2.7) and presentation (7.2.8) of the results of such analyses, whilst the remaining sections of this methodology section outline potential issues in assessing the significance (7.2.9) and reproducibility (7.2.10) of the results of MDA studies. Each of these sections also explains how the methodology applied in the present chapter intends to overcome, or at least mitigate, these issues.

7.2.1 Design bias in the selection of text samples

One of the first question to be addressed in the text selection process is: What is the minimum number of texts that needs to be entered in the analysis to obtain robust results? In text-linguistic research designs, such as those applying MDA, the sample unit is the text, hence the sample size represents the total number of texts. As a general rule of thumb, factor analysis is said to require a dataset of at least five times as many observations (i.e., here, texts) as independent variables (i.e., linguistic features) to be included in the analysis (Hair et al. 2019: 133). Thus, when using Biber's original 67 lexico-grammatical features, a minimum of 335 texts are needed to conduct a full MDA. That said, a high ratio of number of texts to independent variables is desirable (Hair et al. 2019: 133; in the context of MDA specifically, see Friginal & Hardy 2014: 304). As will become evident in the following, the text to linguistic feature ratios in the analyses carried out in the present chapter are, indeed, considerably higher than five.

Section 6.2.2 described how the text units of the Textbook English Corpus (TEC) were defined for the purposes of the additive MDAs carried out in Chapter 6, as well as the steps undertaken to deal with texts that did not reach the chosen minimum text length threshold of 400 words. As compared to Chapter 6, minor changes were made to the text unit subdivisions of the Conversation subcorpus of the TEC (Textbook Conversation). First, as noted in 6.5, the names or character denominations of the interlocutors in textbook dialogues (e.g., *teacher, pupil 1, journalist*) were removed from all the conversation texts of the TEC when they are printed at the start of every line. As part of this semi-automatic process, the subdivisions of the text units of Textbook Conversation were revised to account for the fact that sometimes more than one text is featured on a single CD track or DVD unit. The first process decreased the total number of words in the Textbook Conversation subcorpus from 512,587 to 505,147 whilst the second increased the total number of texts to 593.

TEC texts categorised as Poetry & Rhyme were too few to be meaningfully included in the analyses presented in this chapter; hence, the following multi-feature/multi-dimensional analyses of Textbook English compare the linguistic characteristics of 1,977 TEC texts categorised as belonging to the following five registers: Conversation, Informative texts, Instructional texts, Fiction, and Personal correspondence (see Table 62 for an overview of the TEC texts entered in the present analyses).

Table 62: Textbook English Corpus (TEC) texts processed in this chapter

Textbook Registers	Number of texts	Number of words[46]
Conversation	593	505,147
Fiction	285	241,512
Informative texts	364	304,695
Instructional texts	647	585,049
Personal correspondence	88	69,570
Total	**1,977**	**1,705,973**

Issues pertaining to the selection of the texts or text samples, in particular concerning the comparability of the textbook texts and those of the target language reference corpora, have already been addressed in 3.3. This chapter relies on the same versions of the Youth Fiction and Info Teens corpora used in the previous chapters (see 3.3.2 for details). The only difference between the reference texts tagged for the additive MDA in Chapter 6 with the MAT (Nini 2014; 2019) and those tagged with the Multi-Feature Tagger of English (hereafter MFTE; Le Foll 2021a; 2021e) in the present chapter is that, for the present analyses, the texts of the Spoken BNC2014 included full stops at utterance boundaries to boost the reliability of the feature counts of the tagger (see Online Appendix 3.3 for details of the procedure and Le Foll 2021b for an evaluation of the accuracy of the MFTE on the Spoken BNC2014 with punctuation marks at utterance boundaries).

7.2.2 Design bias in the choice and operationalisation of features

Of all the points of criticism listed at the beginning of 7.2, this one is perhaps the one with the most far-reaching consequences. It was already explained in an early and particularly insightful review of Biber's 1988 publication:

> It is obvious that a method that hinges on statistically determined patterns of co-occurring features will be very sensitive to the selection and identification of these features. If the features are ill-defined, functionally heterogeneous, stylistically skewed, etc., this is likely to have an immediate effect on the results (Altenberg 1989: 171).

In Biber's original MDA design, the linguistic features were chosen on the basis of previous literature (Biber 1988: 71–72). Although Biber and others have always made clear that the range of features entered in MDAs ought to be as broad as possible so as to have the potential to unearth hitherto unseen patterns of variation (Conrad & Biber 2013: 15; Egbert & Staples 2019: 127), such an approach nevertheless risks introducing biases (Diwersy, Evert & Neumann 2014: 174). Altenberg (1989: 173)

[46] As counted by the MFTE.

best illustrates this risk with what he calls the "stylistic 'predisposition'" of some of Biber's (1988) categories. For instance, since they both have connective functions in discourse, Altenberg argues that Biber's (1988) 'conjuncts' category (which includes *alternatively, consequently, further, hence, however*, etc.) is functionally equivalent to the category of 'discourse markers' (which includes *anyway(s), anyhow, now, well*, etc.) and that the distinction between the two is merely situational – the first being specific to literate genres, whilst the second is typical of spoken interactions (see also Siepmann 2004).

To counter the risk of circularity that arises from using top-down feature selection methods, some (foremost computational linguistic) studies have opted for bottom-up approaches to generate features. These approaches correspond to what has also been referred to by some as "corpus-driven research" (see, e.g., Meunier & Reppen 2015: 499; Xiao 2009: 993–996). Such research rejects "*prima facie* those theories, axioms and precepts that were formulated before corpus data became available" (Tognini-Bonelli 2001: 179), hereby avoiding all types of linguistic preconceptions, including those concerning lexico-grammatical categories, e.g., parts-of-speech. Arguably the most data- or corpus-driven approaches to the selection of features involve character n-grams[47]. However, studies based on character n-grams frequently reveal relatively trivial topic-related patterns rather than more generalisable linguistic ones (Baroni & Bernardini 2005: 264; Popescu 2011: 638; Volansky, Ordan & Wintner 2015: 111). Additionally, character n-grams can be argued to lack "direct linguistic motivation or interpretation" (Argamon 2019: 111) and, like token/word n-gram-based methods, they largely fail to account for polysemy. All other data-driven, bottom-up approaches inevitably involve some form of theory-dependent pre-processing steps such as tokenisation, lemmatisation, part-of-speech tagging, (shallow) syntactic parsing, or combinations thereof. It can be argued that such tools add layers of biases in that they rely on specific, pre-established theoretical models of language analysis (cf. Sinclair 1992: 385–390).

In sum, seeking to entirely eliminate bias in the feature selection and operationalisation process whilst nevertheless arriving at a linguistically meaningful and generalisable set of linguistic features may be an unattainable objective. Taking a more optimistic stance, McEnery & Hardie (2011: 114) suggest that bias in the selection of features for MDA can be reduced by ensuring that the selection is "both *principled* and *exhaustive*". The present multi-feature analyses rely on the feature portfolio of a new lexico-grammatical tagger specifically designed for the analysis of situational variation in general spoken and written English: the Multi-Feature Tagger

[47] The first three character tri-grams of this footnote are: *the he_* and *e_f*. Note that in many computational linguistic studies relying on character n-grams, however, whitespaces and/or cross-token n-grams are ignored. When adopting the latter option, the method is arguably no longer truly data-driven since it requires an additional layer of tokenisation.

of English (hereafter: MFTE; Le Foll 2021e). Whilst the MFTE makes no claim to have an entirely "*principled* and *exhaustive*" feature portfolio, numerous steps were undertaken to reduce bias in both the selection and operationalisation of the features. To this end, simplified Hallidayian system networks were examined to ensure that no major aspect of English lexicogrammar would be overlooked (for details, see Le Foll 2021e, cf. Matthiessen 2019; and Whitelaw & Argamon 2004). Ultimately, however, the final choice of features was necessarily restricted by both practical and computational constraints. In particular, the large number of texts usually entered in MDAs means that only features that can relatively reliably be retrieved using automated queries were ultimately included in the feature portfolio of the MFTE. To cite but one example, these constraints resulted in a tagger that makes no distinction between *that*-relative clauses and other *that*-subordinate clauses (unlike the Biber Tagger, whose output, however, is often manual "fix-tagged" for such problematic features; see, e.g., Gray 2019).

Crucially for the interpretability of the dimensions that emerge from the present multi-feature analyses, the MFTE was developed with the aim of arriving at a set of features that not only can be identified to a high degree of accuracy in a variety of written and spoken registers of English, but which can also be meaningfully interpreted in terms of their function. In other words, the developer's aim was that each feature's "scale and values represent[s] a real-world language phenomenon that can be understood and explained" (Egbert, Larsson & Biber 2020: 24). Whilst no automatic tool can ever pretend to be able to achieve this perfectly, countless tagger development-evaluation cycles were completed to arrive at a set of algorithms that best fulfils this criterion. The manifold decisions involved in the selection and operationalisations of these features are detailed in the MFTE's user documentation (Le Foll 2021e; see also Online Appendix 7.1).

Appendix I (see also Online Appendix 7.1) provides a comprehensive list of the final set of over 80 features of the MFTE feature portfolio for which the texts of the TEC and the three reference corpora were tagged for the present analyses (see also Table 63 for an extract). Note that, although the table in Appendix I is subdivided into broad linguistic categories (see also the first column of Table 63 for illustration purposes), these merely serve organisational purposes and do not seek to represent any specific theoretical or functional categorisation. Indeed, many features could equally well be subsumed under a different category. The second column (see also Table 63) provides a very brief description of each linguistic feature. The third corresponds to the tags assigned by the MFTE. Note that these same abbreviations are also used in the tables and figures presented in the results section (7.3). Examples of different language patterns exemplifying these features are found in the fourth column. Finally, the operationalisation column contains simplified, written-out explanations of the combinations of regular expressions used to identify each feature.

For more details, the interested reader is invited to examine the tagger source code available on GitHub (https://github.com/elenlefoll/MultiFeatureTaggerEnglish).

Table 63: Excerpt of Appendix I: Operationalisation of the 'DO as an auxiliary' feature

Category	Feature	Code	Examples	Operationalisation
Verb semantics	*DO auxiliary*	*DOAUX*	*Should take longer than it does. Ah you did. She needed that house, didn't she? You don't really pay much attention, do you? Who did not already love him.*	Assigned to *do, does* and *did* as verbs in the following patterns: (a) when the next but one token is a base form verb (VB) (e.g., *did it work?, didn't hurt?)*; (b) when the next but two token (+3) is a base form verb (VB) (e.g., *didn't it work*); (c) when it is immediately followed by an end-of-sentence punctuation mark (e.g., *you did?*); (d) when it is followed by a personal pronoun (PRP) or *not* or *n't* (XX0) and an end-of-sentence punctuation mark (e.g., *do you? He didn't!)*; (e) when it is followed by *not* or *n't* (XX0) and a personal pronoun (PRP) (e.g., *didn't you?*); (f) when it is followed by a personal pronoun followed by any token and then a question mark (e.g., *did you really? did you not?*); (g) when it is preceded by a WH question word. Additionally, all instances of *DO* immediately preceded by *to* as an infinitive marker (TO) are excluded from this tag.

The MFTE performs feature extraction over several iterations over the texts of a corpus. First, each text is tagged for part-of-speech with the Stanford Tagger (bidirectional version 3.9.2; Toutanova & Manning 2000; Toutanova et al. 2003). Next, rule-based algorithms are run to identify linguistic features necessary for the identification of other features; e.g. DO auxiliaries are first identified on the basis of various combinations of POS tags and forms of the verb DO, before imperatives can be tagged. This ensures that imperative forms of the verb DO can be disambiguated from auxiliary forms, in particular those included in *yes/no* questions where the *do/does/did* frequently occur after an end-of-sentence punctuation mark (see Online Appendix 7.1 for details). Since the Stanford Tagger provides the first layer of linguistic annotation (tokenisation and POS tagging), the accuracy of the feature extraction is heavily dependent on the accuracy of the Stanford Tagger. Whilst it is a well-tested and robust model, it is by no means perfect (Toutanova et al. 2003; Spoustová et al. 2009; Manning 2011). As a result, some of the feature operationalisations outlined in Appendix I include more tags and/or loops then would be necessary if the POS-tagging process were failproof. For instance, since the Stanford Tagger was found to frequently fail to differentiate between past tense (VBD) and past participle forms (VBN), the algorithms designed to capture passives (PASS and PGET) and the perfect aspect (PEAS) include syntactic patterns with

either the VBN or the VBD tag in order to improve recall rates whenever past participles have been erroneously tagged as VBD. Whilst using a POS-tagger as the basis for the feature extraction process reduces the reproducibility of the method as different tagging software (and versions) will inevitably produce different results (Bohmann 2017: 165), the gain in recall and precision is huge and many of the linguistic features of the MFTE's feature portfolio simply cannot be extracted without this initial annotation layer.

Although the MFTE was designed as an all-purpose tagger of general English, its first intended use was for the present project. As a result, some of the feature operationalisations could be adapted to the specificities of the corpora under study. An example of such tailoring concerns the operationalisation of the imperative verb feature. To begin with, the MFTE assigns the imperative tag (VIMP) to tokens identified by the Stanford Tagger as base form verbs (VB), which have not previously been tagged as DO auxiliaries (DOAUX) and are immediately preceded by a punctuation mark other than a comma or such a punctuation mark and an adverb. Textbook instructions often begin with a verb in the imperative; however, these are not always proceeded by an end-of-sentence punctuation mark. Instead, tasks are frequently delimited by a symbol or icon of some kind. These frequently cause OCR issues and produce tokens which are inconsistently identified by the Stanford Tagger as symbols (SYM), list markers (LS) or foreign words (FW). Consequently, the MFTE was designed to also assign to the imperative variable base form verbs which occur after such tokens. It was also noticed that the Stanford Tagger often considers sentence-initial *please* to be a base verb form, hence exceptions were added for the tokens *please* and *thank*. Having identified these sentence-initial imperatives, a second loop then searches for a potential second imperative verb which may occur after *and* or *or* with up to two optional intervening tokens, e.g., (214). Finally, it was noticed that *work* in the phrase *work in pairs*, which occurs more than 700 times in the TEC, is almost invariably identified by the Stanford Tagger as a noun (NN). As a result, this phrase, together with several other frequently occurring phrases which also proved problematic for the POS-tagger were hard-coded as additional exceptions in the version of the MFTE used for the present analyses (version 3.1 ran on perl v.5.22.1 built for x86_64-linux-gnu-thread-multi).

(214) **Describe** or **draw**
Listen carefully and **repeat**
Read the text and **answer** the questions

7.2.3 Uncertainty over the reliability of the feature counts

It is obvious that the robustness of a statistical method that relies on counts of features depends on the high accuracy of these counts; however, very few MDA studies include thorough evaluations of tagger accuracy (two major exceptions deserving of a mention are Biber & Gray 2013; Gray 2015). Crucially, when reporting

the accuracy of a tagger to be used in MDA, it is important to consider not just the overall accuracy of a tagger, since this will be heavily skewed towards very frequent tags – many of which are particularly easy to tag, e.g., punctuation marks and determiners, but also the tagger's accuracy per feature. To complicate matters, accuracy can be measured in various ways (see Table 64). Commonly, only 'precision', i.e., the percentage of correctly assigned tags within a category, is reported. This is for practical reasons since precision is much easier to spot-check than 'recall', i.e., the percentage of a particular feature that is correctly identified as such by the tagger. In practice, however, both precision and recall are important for the results of MDAs to be reliable.

Table 64: Summary of the terminology used in tagger performance evaluation

Term	Definition
True positive	Feature correctly tagged by the MFTE as X
False positive	Feature incorrectly tagged by the MFTE as X
False Negative	Feature incorrectly not tagged by the MFTE as X
Precision	True positive count / (true positive count + false positive count)
Recall	True positive count / (true positive count + false negative count)
F1 score	2 * (precision * recall) / (precision + recall)

The present MDA relies on the MFTE, whose documentation (Le Foll 2021e) details a full evaluation of the tagger's accuracy, including breakdowns per subcorpus of the evaluation corpus (the BNC2014 Baby+) and a table listing recall, precision and the combined F1 measure (see Table 64) for each feature. Whilst the overall accuracy of the MFTE is high and the vast majority of features score above 90% on all three accuracy metrics, the results of the tagger evaluation remain relevant to the interpretation of the results presented in this chapter. Indeed, any potential systematic mismatches between the linguistic constructs under study and the features as they are counted in practice (here, by the MFTE) have the potential to skew any model that emerges from an MDA conducted on the basis of these counts. Picoral et al. (2021) demonstrated that differences of more than 10% in the tagging accuracy of two automatic parsers (the Malt Parser and the Stanford Dependency Parser) and one tagger (the Biber Tagger) are not unusual, yet even such relatively small differences can be of relevance to our understanding and modelling of Textbook English as a variety of English.

Consequently, it was important to ensure that the MFTE performs at least as well on the data used in the present study as on the BNC2014 Baby+. In order to investigate any such potentially problematic mismatches, the results of the tagging procedure were manually checked in six texts from the TEC, originating from six different textbook series (two per country of use). These texts were randomly sampled so as to have one text representing each of the six textbook registers of the TEC (see 3.3.1.4). In total, 4,515 tags were manually checked across these six texts, of

which 114 were found to be incorrect. Excluding punctuation tokens and symbols which were always accurately tagged but whose counts are not entered in the analysis, this means that the MFTE has a satisfactorily high rate of accuracy 97.11% [95% CI: 96.53–97.61%] on the TEC data.

As for the three reference corpora used in the present study, the accuracy of the MFTE on the Spoken BNC2014 data was already evaluated in Le Foll (2021e). Excluding unclear tokens which human annotators could not reasonably interpret from the transcripts (n = 7) and punctuation marks, 5,388 tags were found to be correctly assigned by the MFTE whilst 224 were flagged as incorrect, corresponding to an overall accuracy of 96.01% [95% CI: 95.46–96.51%]. No formal evaluation of the MFTE output was carried out for the remaining two reference corpora because they were deemed to be very similar to subcorpora of the BNC2014 Baby+, for which a thorough investigation has already been conducted. The Youth Fiction corpus is highly comparable to the fiction subcorpus of the BNC2014 Baby+, for which the MFTE accuracy was found to be very high (96.93% [95% CI: 96.46–97.39%], excluding punctuation and symbols), whilst the Info Teens corpus shares similarities with the news subcorpus and some of the e-language subregisters, in particular: blogs and forum posts (see Brezina, Hawtin & McEnery 2021). The evaluation files corresponding to these (sub)registers of the BNC2014 Baby+ from the original tagger evaluation (Le Foll 2021e) were thus reanalysed for the purposes of the present study. On these eight news articles, blog, and forum posts, the MFTE reached a slightly lower, but nevertheless satisfactory, accuracy rate of 95.84% [95% CI: 95.36–96.28%] excluding punctuation and symbols.

Taken together, the overall MFTE accuracy for the TEC and the three reference corpora, excluding unclear tokens, punctuation, and symbols, can thus be estimated to be around 96.38% [95% CI: 96.13–96.62%]. The per-feature accuracy measures are visualised in Fig. 49. The full breakdown of the evaluation results and the corresponding code can be found in the Online Appendix 7.3.

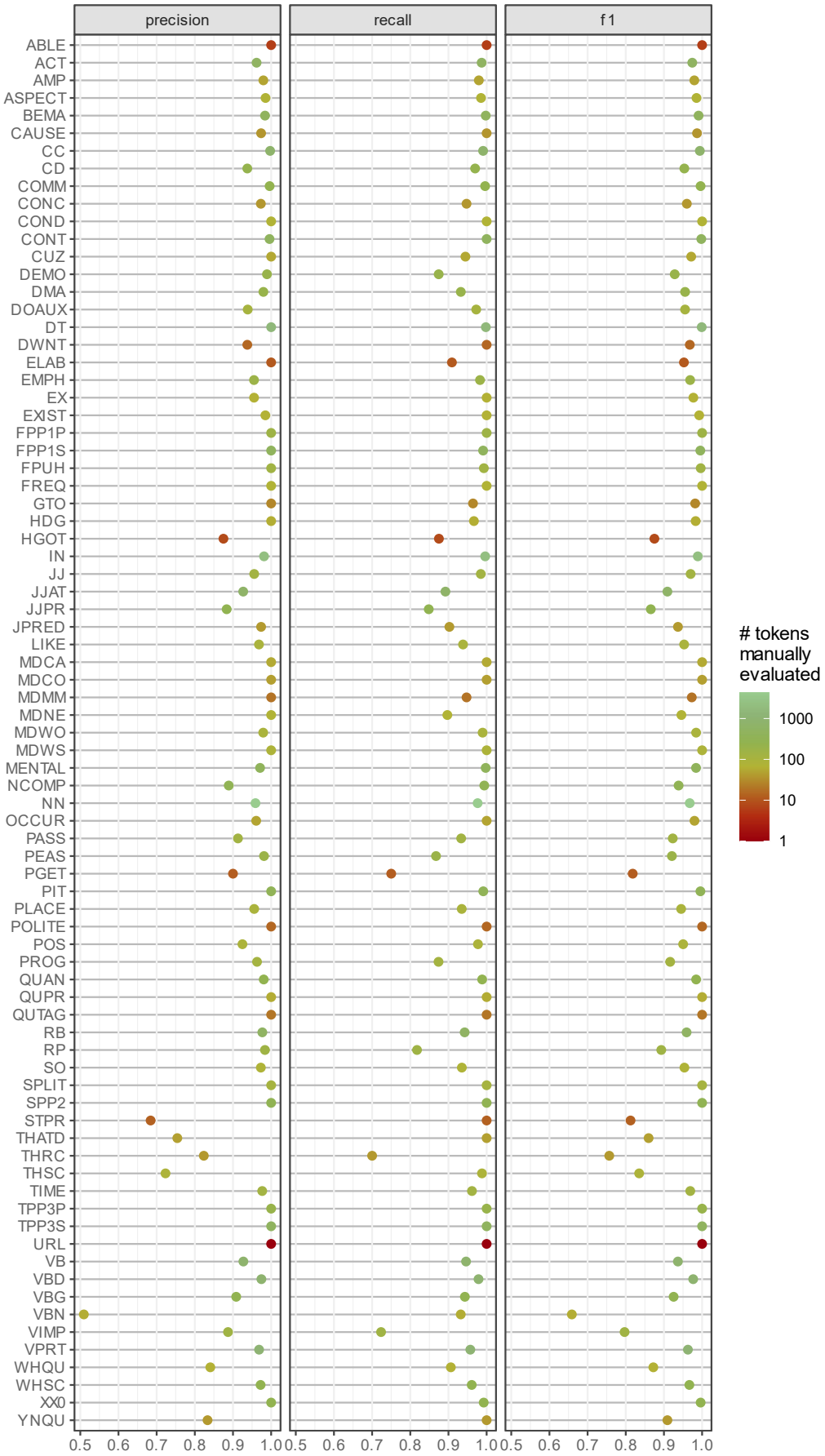


Fig. 49: Per-feature accuracy measures of the MFTE on the TEC, the Spoken BNC2014 and data comparable to the Youth Fiction and Info Teens

In total, 198 different error types (combinations of an erroneous tag assigned by the MFTE and of the corrected tag) were identified as part of the present tagger evaluation. Nearly half of the error types (n = 94) were only identified once. Table 65 lists the most frequent error types which were found 12 or more times across the six evaluation texts of the TEC and the BNC2014 Baby+.

Table 65: Most frequent tagging error types

Error type (MFTE tag -> corrected tag)	N	Example of error type (incorrectly tagged token and tag in bold)
NCOMP -> NULL	37	my part-time_NN **book_NN** **NCOMP** - stacker
NN -> JJAT	35	a six-game **losing_NN** run
JJAT -> NN	27	it was the **backroom_JJAT** staff
NN -> VB	27	YOU MUSTN'T: **chat_NN** whilst your classmate
IN -> RP	25	He could put **on_IN** a bit of a show
NN -> VPRT	24	Suzannah the tail-wagger **wags_NN** the tail of
VB -> NN	22	would expect them to deliver their first **win_VB**
THSC -> DEMO	19	I'll do **that_THSC** a bit later cuz
VB -> VIMP	19	List events and **place_VB** them on a timeline
NN -> OCR	16	Be the most creative! Thin_JJ **k_NN**: Why?
VBN -> JJAT	16	a barely **published_VBN** novelist
ACT -> NULL	15	guilty of a foul **throw_VB** **ACT** in his own half
THATD -> NULL	15	that's how you **remember**_VPRT **THATD** things isn't it?
CD -> NN	12	there's a good **one_CD**
MENTAL -> NULL	12	No there's loads in here I **like_VPRT** **MENTAL** there's a good one

A qualitative analysis of the most frequent tagging errors showed that the first source of errors arose from the texts themselves: the MFTE most frequently mistagged tokens that were either nonsensical as the result of OCR errors or typos/non-standard spellings in the original texts.[48] In spite of intensive manual corrections, the scanned textbooks of the TEC, in particular, still contain many OCR errors, whilst the e-language subcorpus of the BNC2014 Baby+ features many typos and non-standard spellings (and, indeed, this is also the case in the Info Teens). Unrecognised words are usually assumed by the Stanford Tagger to be nouns, which is why many of the most frequent error types listed in Table 65 involve the MFTE mistagging tokens as nouns (NN) and identifying noun-noun compounds (NCOMP) when there are none. In such instances, the human annotators added the NULL tag in the correction field (for details, see Le Foll 2021e).

[48] Note that, for the purposes of the present evaluation, tokens that were poorly processed but nevertheless correctly tagged were not counted as tagging errors (e.g., when *Helen* was misrecognised as *Llelen* by the OCR software but still correctly identified as a noun by the MFTE).

The second major source of tagging errors stems from the underlying layer of POS-tagging carried out by the Stanford Tagger that the MFTE relies on for nearly all its algorithms. These errors typically involve confusing the POS of word forms with more than one possible POS. For example, in one of the evaluation files, the word *build* occurred four times as a noun in the compound *beta build* but was erroneously identified as a verb by the Stanford Tagger. As explained in 7.2.2, the MFTE was designed to counter some of the most systematic frequent errors that the Stanford tagger was found to make (e.g., identifying the *'s* in *let's* as a possessive marker, or better differentiating between past participle and past tense forms). Nevertheless, when a word form with different possible POS is found in an unusual context, the most frequent of the possible POS is usually assumed by the Stanford Tagger.

Finally, the third largest source of feature identification errors is due to how the features are operationalised in the MFTE. For many of the more complex grammatical features, in particular, the algorithms of the MFTE outlined in Appendix I and in the Online Appendix 7.1 are best approximations, which are designed to capture the majority of these features, but they are by no means expected to be 100% accurate in all contexts. The tagging issues related to the incorrect identification of imperatives (VIMP), verbs of action (ACT), and *that*-omissions (THAD), in particular, belong to this category.

A full list of all the erroneously assigned tags and their corrections can be found in Appendix 7.3. They will not be discussed any further here because they are very similar to those found in the formal evaluation of the MFTE carried out on the BNC2014 Baby+ data (Le Foll 2021e). In sum, the results of this brief evaluation confirm that the corpora used in the present study are suitable for tagging with the MFTE.

7.2.4 Confusion between covariation of features due to situational variation and covariation due to grammatical structure

Most MDAs are carried out on collections of texts that vary in length. However, counts of linguistic features in texts of different lengths cannot be directly compared. Let us imagine a short business e-mail of 200 words that features four occurrences of the word *if* and compare this to an imaginary novel of, say, 20,000 words in which the word *if* is observed eight times across the entire book. Comparing these raw counts, we may naively be tempted to conclude that *if* is twice as frequent in fictional writing as in professional e-mails. Evidently, however, this comparison does not account for the vastly different number of potential opportunities of use of the word *if* in the two texts. To remedy this, the *de facto* standard in corpus linguistics has so far consisted in normalising raw counts to a common word-based denominator. For example, the count of four *if*s in our hypothetical e-mail can be divided by the total number of words in the e-mail (200) and the count of eight in the book by the length

of the novel (20,000) before multiplying both results by a common denominator, e.g., by 1,000 words. In our example, this approach results in normalised *if* frequencies of 20 per 1,000 words in the e-mail and 0.4 per 1,000 words in the novel. In other words, once normalised on a per-word basis, we might conclude that *if* is in fact 50 times more frequent in professional e-mails than in novels! Typical word-based denominators are 100, 1,000, 10,000 and a million words and the resulting numbers are referred to as normalised, normed, or relative frequencies.

Word-based normalisation baselines imply that words are used independently of each other and therefore do not account for, or attempt to model, the actual choices that language users make when producing (or, for that matter, when processing) language. Thus, word-based normalisation baselines can be said to conflate frequency of use and opportunity of use (Wallis 2020: 47–52). In reality, once a language user has chosen a particular word, they have a limited number of choices at their disposal as to which word can logically come next. This phenomenon is illustrated in Fig. 50. It displays the most frequent word forms that immediately follow the word *if* in the Spoken BNC2014. It shows that more than a third of occurrences of *if* are followed by the word *you* and that the seven most frequent word forms after *if* are all personal pronouns. Thus, Fig. 50 makes clear that not all words (or word classes) are equally likely to occur after the word *if*.

		Word	Frequency ↓	Relative ?	% of conc. ?		
1	☐	you	18,299	1,546.45	38.70%		•••
2	☐	I	6,520	551.00	13.79%		•••
3	☐	it	4,303	363.65	9.10%		•••
4	☐	they	3,429	289.78	7.25%		•••
5	☐	we	2,450	207.05	5.18%		•••
6	☐	he	1,905	160.99	4.03%		•••
7	☐	she	1,209	102.17	2.56%		•••
8	☐	that	1,022	86.37	2.16%		•••
9	☐	there	1,016	85.86	2.15%		•••
10	☐	if	980	82.82	2.07%		•••

Fig. 50: The ten most frequent word forms occurring immediately after *if* in the Spoken BNC2014 (as counted and displayed by Sketch Engine, see https://ske.li/nu8 for full results)

Corpus linguistics has a long history of using word-based normalisation rates indiscriminately and the MDA framework is no exception. Whilst per-word normalised frequencies can be argued to represent language users' rates of exposure (which may well be what some researchers, e.g., lexicographers, are attempting to model), a number of recent publications have pointed out potential issues with the indiscriminate application of this approach. Wallis (2020: 56), for instance, explains

how per-word frequencies undermine the assumptions of many of the statistical models used in corpus linguistics which assume that linguistic features follow binomial distributions, i.e., that it is, in principle, possible to observe proportions of 100%. In practice, however, it is highly improbable that a language user would simply repeat a single word or use only words of a single word class for the entire duration of a text! In the Biberian MDA framework, the indiscriminate use of word-based normalisation for all linguistic features results in a highly problematic confusion between the covariation of features due to situational variation (which of course is what the method actually aims to tease out) and covariation due to grammatical structure (see also Grieve-Smith 2007). Consider the linguistic features that load on Biber's (1988) Dimension 1: at the positive, 'involved' end, these include the number of verbal contractions, negated verbs, and present tense verbs per 1,000 words, which all correlate strongly with each other. Whilst it is true that these features are particularly frequent in spoken interactions, it is also undeniable that these correlations are inevitably mediated by the overall frequency of verbs. Similarly, the high positive correlations that contribute to negative scores on Biber's (1988) Dimension 1, e.g., between the per-1,000-word normalised frequencies of nominalisations, determiners and, though to a slightly lesser extent, prepositions, are all "grammatically mediated" by the frequency of nouns.

In the context of MDA, the choice of the normalisation units for the linguistic features to be entered in a model is anything but trivial. For a start, factor analysis (FA) and principal component analysis (PCA) are both "exquisitely sensitive to the sizes of correlations" (Tabachnick & Fidell 2014: 665). Consequently, the authors warn that "it is critical that honest correlations be employed" (Tabachnick & Fidell 2014: 665, cf. 95–96). However, covariation due to grammatical structure can be argued to generate linguistically "obvious correlations" (Evert 2018: 24) and, as such, "dishonest" ones.

That being true, the choice of normalisation unit(s) for the counts to be entered in MDAs is by no means a simple one. Reflecting on the quantification of linguistic measures in general terms, Schegloff (1993: 103) speaks of the need to account for "environments of possible *relevant* occurrence" (emphasis original) and argues that "quantitative analysis requires an analytically defensible notion of the denominator". In practice, however, the choice of the denominator of normalised frequencies, the normalisation baseline, will depend on both the linguistic conceptualisation of a linguistic feature, as well as the feasibility of reliably counting what is considered to be the most meaningful unit to capture an opportunity of use (cf. Wallis 2020: 69–70). For instance, as a denominator for the counts of *if*, the total number of sentences or clauses in a given text might seem like the most linguistically meaningful or "analytically defensible" unit. However, whilst identifying sentences is relatively trivial in written registers, not only can this unit be argued to not make much

linguistic sense in spoken registers[49], but it is also impossible to reliably implement with spoken corpora whose transcription scheme do not include any sentence boundaries (e.g., the Spoken BNC2014). As for automatically identifying clauses, this would require dependency parsing, which, to date, remains highly unreliable for transcriptions of spontaneous spoken language, and which would, in any case, certainly result in units that would be equally difficult to compare across different modes and registers.

The present analyses rely on the normalisation baselines as implemented in the "complex normalisation" output of the MFTE (see Le Foll 2021e for details). For this output, the MFTE normalises counts of the majority of features including conditional conjunctions (*if* and *unless*; COND), contractions (CONT), negation (XX0), present tense verbs (VPRT), and WH-questions (WHQU) to 100 finite verb phrases. The number of finite verb phrases is approximated to a satisfactorily high degree of accuracy by the MFTE by adding the counts for present tense (VPRT), past tense (VBD), imperatives (VIMP) and all the modal verbs (MDCAN, MDCOU, MDMM, MDNE, MDWO, MDWS) together. Five features, attribute adjectives (JJAT), *s*-genitives (POS), noun compounds (NCOMP), quantifiers (QUAN), and determiners (DT), are normalised to 100 nouns, whilst only the remaining 19 features, e.g., emoji and emoticons (EMO), discourse markers (DMA) and nouns (NN), are normalised to 100 words.

7.2.5 Lack of transparency in the quantitative patterns captured by factor analysis

In the Biberian framework, MDA relies on factor analysis to reduce a large set of associations of normalised counts of many different linguistic features across a large number of texts to a more parsimonious set of underlying, or latent, variables. These summarising variables are first referred to as 'factors' and then, once they have been functionally interpreted, as 'dimensions'. Thus, MDA makes use of factor analysis to reduce complexity and "consolidate variables in a principled manner" (Loewen & Gonulal 2015: 183) in order to more concisely describe, and ultimately hopefully understand, the relationships among the linguistic features. The underlying belief is that such parsimonious solutions will have greater external validity and will therefore be more likely to replicate (Henson & Roberts 2006: 394).

At this stage, it should be noted that the terminology is often used ambiguously and that statisticians disagree as to what exactly does or does not constitute 'factor analysis' (see, e.g., Henson & Roberts 2006: 398; Jolliffe 2002: 150). In the present thesis, 'factor analysis' will be used as an overarching term that includes 'common

[49] "A sentence is a constituent of writing, while a clause complex is a constituent of grammar" (Halliday 1993: 216).

factor analysis' and 'principal component analysis' (see explanation below). Following Biber (1988), the factor-extracting method of choice in MDA studies has traditionally been exploratory factor analysis (EFA), which is a common factor analysis method. Although central to the MDA methodology, a meta-analysis of MDA studies published in English or Portuguese between 1984 and April 2020 found that as many as 69% (n = 65) of the examined studies did not report which factor extraction method they used (Goulart & Wood 2021: 124). Of those that did, 78% (n = 25) reported using EFA.

The use of EFA as a means of reducing complexity in MDA studies has been criticised as "lacking in transparency" (Evert 2018: 12). One of the reasons for this is that the results of EFAs are contingent on the number of factors retained by the researcher(s). This means that solutions in EFA are not unique, but rather many different dimension scores can be computed for each observation (Lee 2000: 171). This potential for researcher bias is well-documented which is presumably why all best-practice guidelines on how to conduct and report EFAs (e.g., Tabachnick & Fidell 2014: 696–699; Loewen & Gonulal 2015: 194–197) devote a section to various (more or less objective) factor retention criteria. The problem is that, whilst methods to determine the number of factors to retain abound (e.g., Kaiser-1 rule, Joliffe's criteria, visual inspection of the scree plot, parallel analysis…), they each tend to produce vastly different results. Crucially, factor indeterminacy in common factor analysis means that the results of EFAs are not computationally stable. In other words, using the same data, researchers will obtain different results for, say, the first three factors of an EFA model depending on whether they decide to extract three, four, five, or more factors.

Within the traditional MDA framework, determining the number of factors to extract has been described as an "important part of the iterative process between statistical procedure and subjective researcher interpretation" (Egbert & Staples 2019: 130). Comparing the interpretability of the various combinations of factors that emerge from different factor solutions is considered a valid approach. It is argued that:

> If particular factor solutions (greater number or smaller number) are more interpretable than others, then it should be considered as a more favorable solution (Egbert & Staples 2019: 130).

Given that the number of factors to be extracted has a more profound impact on the final solution than any of the other decisions related to the statistical method employed (e.g., threshold levels, EFA vs. PCA, rotation method) (Lee 2000: 99), the risk of (conscious or unconscious) researcher bias that such an approach entails is obvious. This is one of the reasons why Diwersy et al. (2014) and Neumann & Evert (2021) advocate for the use of principal component analysis (PCA) rather than EFA. Whilst researchers conducting PCAs still need to choose one of the many methods to decide on how many summarising, latent variables (referred to as 'principal

components' in PCA) to retain and interpret, the results themselves will remain the same regardless of how many components are deemed to be worthy of further analysis and (linguistic) interpretation (Jolliffe 2002: 159–160).[50] Several multi-feature and multi-dimensional linguistic studies have already been successfully conducted using PCA. Notably, Sigley (1997) convincingly applied PCA to develop a formality index of English, whilst Neumann & Evert (2021; based on a method originally developed in Diwersy, Evert & Neumann 2014) have proposed a new approach inspired by Biber's MDA framework, Geometric Multivariate Analysis (GMA), that uses the orthogonal projections (i.e., those of mathematically independent, uncorrelated, factor axes) of PCA to explore differences between texts and groups of texts in multi-dimensional feature space.

Like common factor analysis, PCA is also a dimension-reduction statistical method. However, mathematically, EFA and PCA differ in that PCA accounts for all the variance in the data, thus pooling together shared (or 'common') variance between variables, specific variance, and error variance, whilst EFA attempts to estimate the variance due to error and the variance that is unique to each variable in order to eliminate these sources of variance to focus exclusively on the shared (or common, hence the term 'common factor analysis') variance. Whilst the latter may produce factor solutions that are more readily interpretable, the computing of the factors themselves is rather opaque. By contrast, the components produced by PCA are linear functions directly derived from the observed variables, i.e., in the present context, from the normalised counts of each linguistic feature in each text of the corpora.

Theoretically, too, EFA and PCA differ in that EFA produces latent variables that are assumed to represent real-life constructs. Conceptually, these constructs are thought to "cause" the distributions of the variables to be as they are observed in the dataset whereas, in PCA, it is the resulting components that are "caused by" or that "produce" the observed variables. Thus, the components of PCA can be said to be empirically real factors that directly represent aggregates of the observed correlated variables. However, these do not necessarily reflect any underlying constructs or processes (Tabachnick & Fidell 2014: 662).

In practice, however, specific and error variance (also known as 'unique variance') are often not large enough to significantly affect the first few components of PCA, so that, when focusing on these first few dimensions, the choice of EFA over PCA will usually lead to the same interpretation of the results (Lee 2000: 168–170; Tabachnick & Fidell 2014: 634). Furthermore, Henson & Roberts (2006: 398) explain that any differences in the outputs of the two methods will "decrease as (a) the measured variables have greater score reliability or (b) the number of variables measured

[50] Note, however, that this is no longer true if the PCA solution is rotated (see, e.g., Husson, Lê & Pagès 2017: 29).

increases." Indeed, preliminary exploratory factor and principal component analyses conducted on the TEC and its three reference corpora as tagged by the MAT produced very comparable results (Le Foll 2020a). However, for the present chapter, PCA was chosen for its computational stability and its (relative) ease of interpretation. The latter is also true of the feature weights (loadings) which, in PCA, simply represent correlation coefficients between the observed variables and the components, whereas the factor loadings that emerge from EFAs are factor score estimations. These are mathematically much more complex and, although they fulfil a very similar function, are therefore more difficult to accurately interpret.

A quest for simplicity is also what motivated the interpretation of true principal components as opposed to rotated components. In applying the MDA framework, most linguists follow Biber (1988: 84–85) and apply a rotation to the extracted factor loadings of their EFA solutions. In brief, such rotations are often recommended because they transform the solution such that the extracted factors (or in PCA: the components) are strongly correlated with certain variables and uncorrelated with others. Thus, the procedure simplifies the factor loading structure which is often thought to make the results more readily interpretable. However, this supposed simplification comes at a cost: such rotated factors or components no longer maximise group separations and the rotated loadings across the factors are correlated (Rencher 1992: 219) – something which Biber and others tend to brush aside in their interpretations of such rotated solutions (see Lee 2000: 253–256). In addition, Biber (1988) chose to apply a Promax 'oblique rotation' which, unlike the more commonly used 'orthogonal rotation' methods, allows for the correlation of factors. Biber (1988: 85fn.) convincingly justifies his choice by explaining that:

> [O]blique solutions might be generally preferable in studies of language use and acquisition, since it is unlikely that orthogonal, uncorrelated factors actually occur as components of the communication process. That is, from a theoretical perspective, all aspects of language use appear to be interrelated to at least some extent, and thus there is no reason to expect mathematically uncorrelated factors representing those aspects (see Hinofotis 1983).

In practice, however, producing oblique solutions considerably complicates the interpretation of the results. Since oblique rotations allow for a degree of correlation between dimensions, both the factor pattern and the factor structure matrices need to be interpreted. However, Biber (1988) and seemingly most linguists who follow Biber in applying oblique rotations only interpret the factor pattern matrix. The degree to which this impacts the results of such studies is difficult to evaluate because very few MDA studies conducted with oblique rotations report the inter-factor correlation coefficients. In the few that do (e.g., Biber 1988: 84; Biber & Egbert 2016: 12), significant inter-factor correlations are reported though these are either not commented on or merely brushed aside as "small" (Biber 1988: 85) or "generally

small" (Biber & Egbert 2016: 12)[51] even when the results actually point to several inter-factor correlations above 0.3 or even 0.4. As Lee (2000: 247–254) demonstrates in his replication of Biber's (1988) seminal MDA, ignoring inter-factor correlations can have a profound impact on the interpretation of the resulting factor solution.

In the case of rotating PCA solutions specifically, another disadvantage of applying rotation is that the results are no longer independent of the number of extracted components. In other words, using the same dataset, the first two components of a two-dimensional rotated solution will not be the same as the first two components of a three-dimensional rotated solution (Husson, Lê & Pagès 2017: 29), which once again raises the million-dollar question of the "correct" number of components to extract (see above).

As a result of all the above considerations, the MDAs conducted in this chapter rely on PCA rather than EFA. Visual inspection of eigenvalues scree plots is used as a first step to determine how many components are to be analysed. To make the most of different visualisation options, two different R packages are used to conduct the PCAs: {stats} for its stats::prcomp function that allows for 3-D visualisation of the results via {pca3d} and {PCAtools} for its highly customisable PCAtools::pairsplot and PCAtools::biplot functions for 2-D graphs (for details of all the packages, functions and parameters used, see code in the Online Appendix 7.2–7.7).

7.2.6 Degradation of correlations

As a family of statistical methods, dimension-reduction methods are known to be very sensitive to outliers and skewed distributions of variables. Tabachnick & Fidell (2014: 665) lament that "problems created by missing data, and degradation of correlations between poorly distributed variables all plague FA [factor analysis] and PCA". These issues, however, are rarely discussed and, up until now, have largely been overlooked in MDA studies.[52]

In the context of MDAs with linguistic features, it is perfectly possible for features to be entirely absent from some of the texts in the corpora under study, thus creating the impression of "missing data" in the count matrices to be entered in such analyses. Of course, the data is not "missing" in the traditional sense but rather the rate of occurrence of these features is simply zero. There are, in theory, three reasons why this might be the case. The first is quite simply that a text genuinely does not feature this particular lexico-grammatical unit. For instance, it is easily conceivable that a

[51] Since Biber & Egbert (2016) conducted a PCA rather than a common factor analysis method such as EFA or CFA, these are in fact inter-component correlations.

[52] As the reader will notice in the following, Lee (2000) constitutes a notable exception; however, his PhD thesis has not yet been published and, as a result, his concerns regarding the use of EFA with untransformed variables appear not to have reached the wider (corpus) linguistic research community.

novel may not include a single emoji or emoticon (EMO). Thus, especially linguistic features that have "strong stylistic discriminating properties" (Lee 2000: 173), such as emojis and emoticons, will necessarily follow very skewed distributions across multi-register (or multi-dialect, multi-variety, etc.) corpora. Moreover, texts as long as an entire novel are rarely entered in an MDA. For example, the Youth Fiction Corpus analysed in the present study consists of extracts of novels of around 5,000 words each. Second, therefore, we may envisage a situation in which a particular feature is absent from a text extract, thus returning a count of zero, but can actually be observed in other parts of the full text. For instance, a short extract of a novel may not happen to include a single verb in the passive voice, yet it is highly unlikely that the entire novel does not feature a single verb in the passive. Similarly, if a complete text is very short it is also likely to have zero occurrences of many of the least frequent linguistic features, even though this may not be representative of the text register/variety more generally. For instance, if we examined a small corpus of tweets, we may find that there are zero occurrences of the word *because* in most tweets. A hasty functional interpretation of this finding could be that Twitter discourse favours statements of facts and opinions rather than explanations. In reality, however, we might find that pooling a random sample of texts as short as tweets from any other register, be it news reports, novels, or speeches, may yield the same result. In other words, in some cases, and especially for relatively rare linguistic features, zero counts can occur simply by virtue of texts being very short. Ultimately, both these issues tie back to the discussion of an appropriate minimum text length for MDA studies (see 6.2.2). Therefore, solutions such as those adopted for the present thesis, i.e., the ordered collation of particularly short texts (particularly relevant for the TEC; see 6.2.2), and the random sampling of extracts of longer texts (see 3.3.2) must be carefully considered in order to avoid such issues causing undue influence on models derived from MDA. Finally, it should also be acknowledged that a third reason why a text may appear to include zero occurrences of a particular feature may be due to a failure of the automatic tagger used to identify and count the feature in question. This risk confirms the need to conduct thorough evaluations of the taggers that are used in MDAs, as discussed in 7.2.3.

Thus, we have seen that there are potentially several reasons why count matrices destined to be entered in MDAs may include zeros. For some of these reasons, mitigating steps have already been undertaken as part of the pre-processing of the corpus data. However, it may still be necessary to remove linguistic features that are genuinely very poorly distributed as a result of being entirely absent from a large proportion of the texts to be analysed using factor analysis. In the present analyses, features with zero occurrences in more than two thirds of texts were therefore either excluded from the analyses (see Bohmann 2017: 168 for a similar procedure) or, whenever linguistically meaningful, merged with other features.

Even having removed features with high percentages of zero occurrences, the distributions of many of the remaining features nevertheless remain highly skewed. By way of illustration, the normalised frequencies of occurrence of five features across the TEC are plotted in Fig. 51. A cursory look at these example histograms points to two potential issues. First, unsurprisingly, the ranges of rates of occurrence (plotted on the *y*-axes) vary considerably. These ranges depend a) on how frequent a particular linguistic feature is (e.g., we would expect nouns [NN] to be generally much more frequent than split auxiliaries [SPLIT]) and b) on each feature's normalisation basis (e.g., here, the normalised counts of nouns represent the number of nouns per 100 words, whereas progressives [PROG] and split auxiliaries are counted per 100 finite verbs). Second, it is obvious that at least three of these distributions are far from normal and, instead, appear to follow distributions sharing similarities with the Zipfian distribution that is very familiar to linguists (see, e.g., Brezina 2018: 44–46).

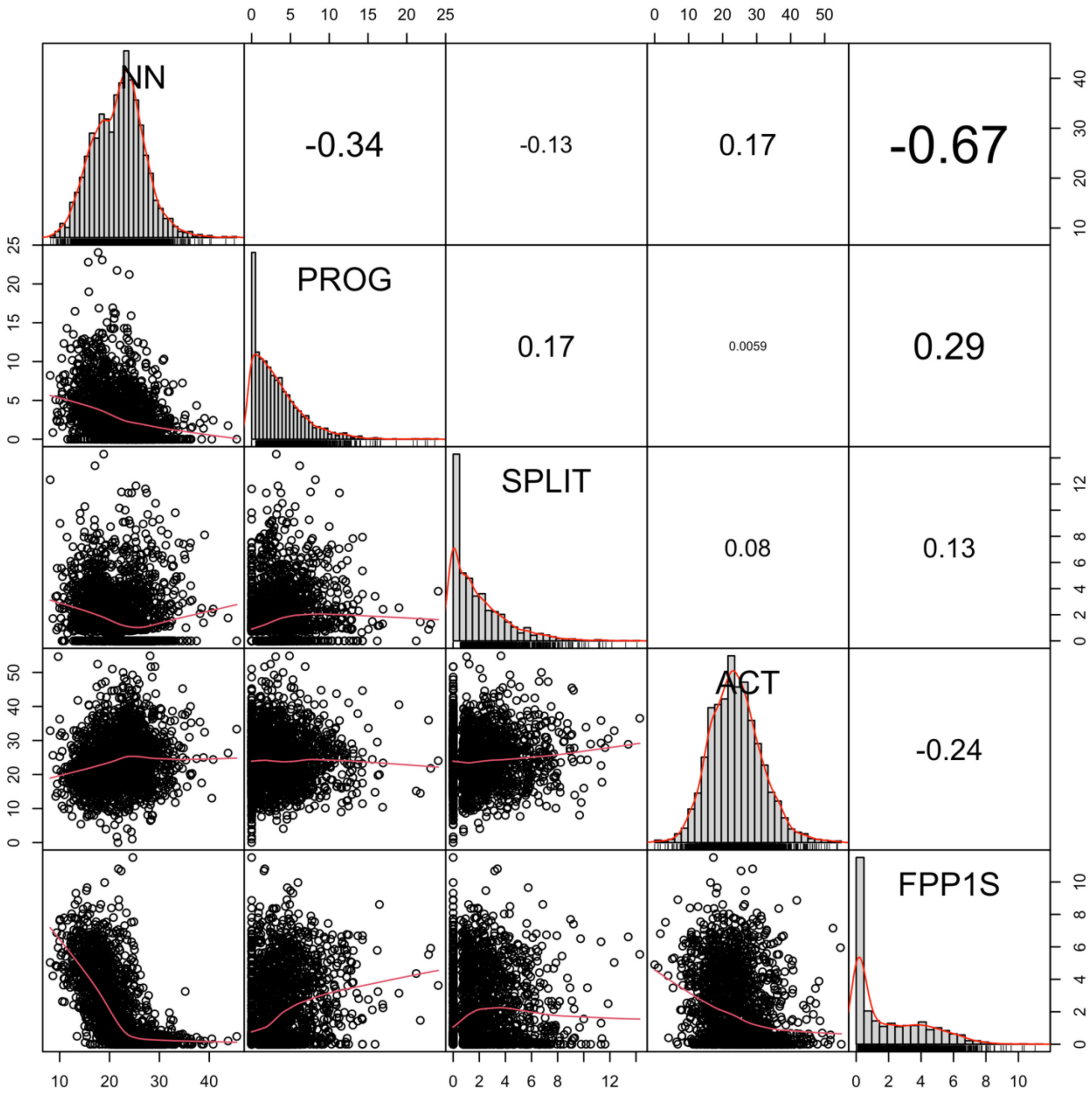


Fig. 51: Distribution of normalised frequencies of occurrence of five features across the TEC (histograms) and visualisations of their correlation (scatterplots)

Dealing with the first issue is relatively trivial: in such cases, it is common practice in multivariable analyses, and indeed in Biber's MDA framework, too (Biber 1988: 94–95), to standardise variables to *z*-scores, i.e., to scale all frequencies to a mean of zero ($\mu = 0$) and a unit variance of one ($\sigma^2 = 1$). This *z*-transformation ensures that each feature makes the same overall contribution to the distances between the texts that will be explored in the following analyses (see Neumann & Evert 2017: 53, see also 6.2.4). As for the second issue, Hair et al. (2019: 137) note on the assumptions of exploratory factor analysis that:

> [f]rom a statistical standpoint, departures from normality, homoscedasticity, and linearity apply only to the extent that they diminish the observed correlations.

In other words, factorial patterns may be harder to detect if variables are not normally distributed and if correlations are nonlinear (as shown in the scatterplots in Fig. 51), but, if/when they *are* detected, there is no reason to assume that they are not real. Nonetheless, as pointed out at the beginning of this section, such skewed distributions run the risk of outliers exerting undue influence on the resulting models. Hence, following Neumann & Evert (2021), the standardised normalised counts were subjected to a signed log transformation[53] in order to (partially) deskew their distributions. The results of these two transformations, *z*-standardisation and log-transformation, are exemplified in Fig. 52. A comparison of the scatterplots in Fig. 51 and Fig. 52 shows that, although these transformations have strengthened the correlations, many of the distributions remain paranormal and the relationships between some pairs of features remain nonlinear. As emphasised by Tabachnick & Fidell (2014: 666), however, as long as PCA (or EFA for that matter) is used descriptively, "assumptions regarding the distributions of variables are not in force." However, they also make clear that when normality and linearity fail, the produced solutions are "degraded" – though this is not to say that they may not "still be worthwhile" (Tabachnick & Fidell 2014: 666–667). Note that, whilst Neumann & Evert (2021) apply a signed logarithmic transformation to deskew feature distributions as an alternative to removing very sparse features, the present methodology uses a combination of methods: removing any features that occur in fewer than a third of texts and transforming the remaining features' standardised normalised counts.

[53] The following function was applied to the standardised normalised counts: signed.log <- function(x) {sign(x)*log(abs(x)+1)} (see Online Appendix 7.4–7.5 for details of the procedure)

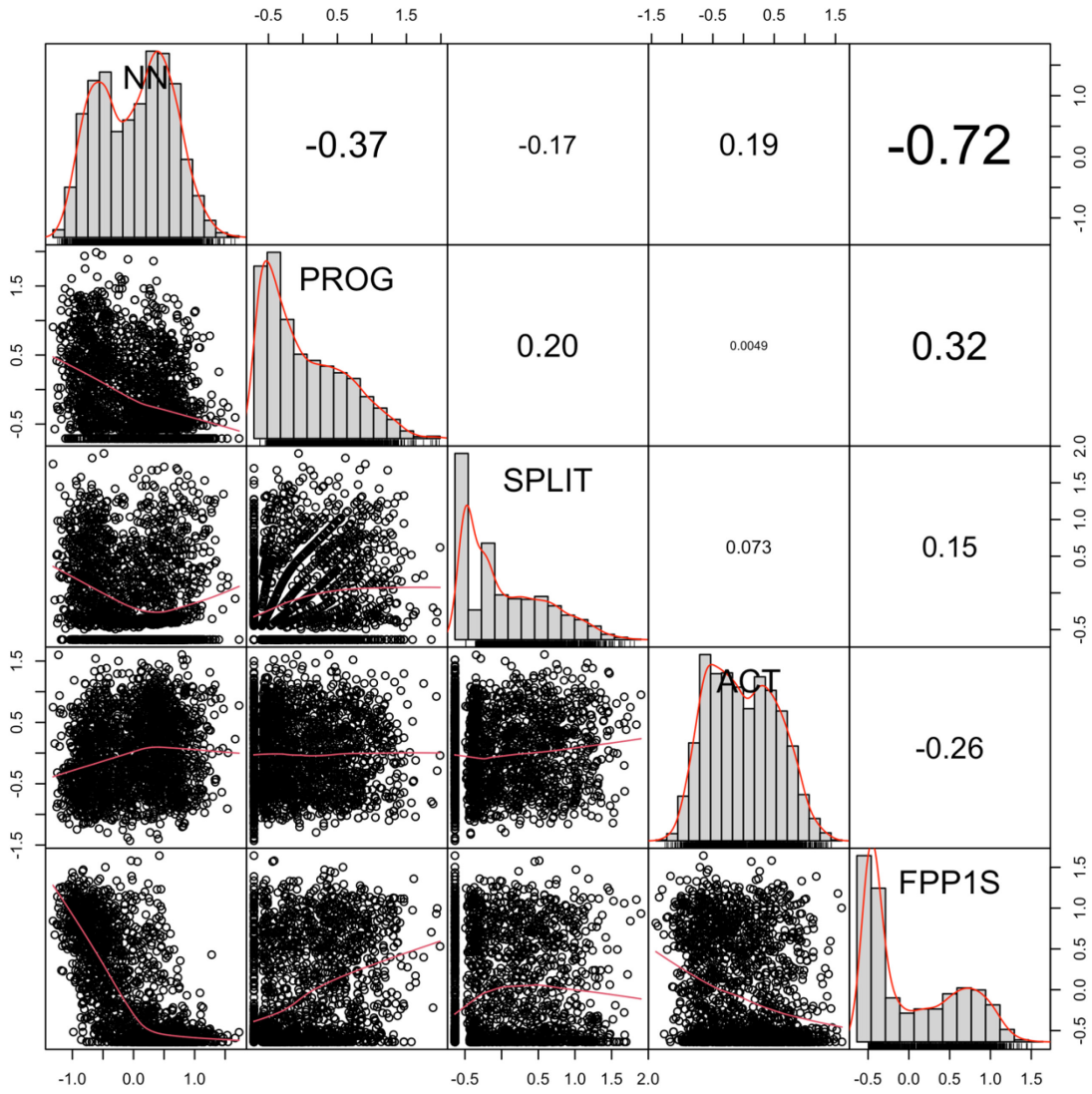


Fig. 52: Distribution of signed log transformed standardised normalised frequencies of occurrence of five features across the TEC (histograms) and visualisations of their correlation (scatterplots)

Overall, the factorability of the data depends on both the number and size of its variable intercorrelations. Bartlett's test of sphericity is often used to test whether variables are sufficiently intercorrelated to produce representative factors; however, it is a significance test of the hypothesis that the correlations in a correlation matrix are zero and, as such, is known to be overly sensitive and dependent on the sample size (Hair et al. 2019: 136) so that, in practice, it will always produce significant results in the context of MDAs carried out on sufficiently large corpora. As formulated in the null hypothesis that it is designed to test, Bartlett's test of sphericity merely indicates the presence of non-zero correlations, which is not to say that the pattern of these correlations is actually suitable for factor analysis (see also Hair et al. 2019: 168). In addition, the test assumes that the data is a sample from a multivariate normal population which is rarely, if ever, the case when dealing with linguistic data (Lee 2000: 178). Thus, for MDA studies with many data points and non-normal

variables, it is wiser to first examine feature intercorrelations visually (see, e.g., Fig. 53, see also Neumann & Evert 2021). This can help to identify both extremely high correlations (collinear variables) and very low ones that can skew the results of the MDA and therefore ought to be excluded. In the present thesis, collinear features are defined as those correlating $> |0.95|$. Whenever this is the case, the less marked of the two collinear variables is excluded from the analysis. For example, in the second MDA presented in this chapter (7.3.2), the present tense (VPRT) and past tense (VBD) variables correlated at -0.96, leading to VPRT being removed from the dataset. In addition, the Kaiser-Meyer-Olkin (KMO) index (1974: 112) is used to further explore the suitability of the feature intercorrelations for factor analysis. In the present thesis, this is achieved using the R psych::KMO function (Revelle 2020) which outputs an overall KMO Measure of Sampling Adequacy (MSA), as well as MSA scores for each individual feature. These can range from 0 (i.e., not in any way correlated with another feature) to 1 (indicating that this feature can be perfectly predicted by another feature) (Kaiser & Rice 1974). Following the procedure described in Hair et al. (2019: 136–137), the features' individual MSA values are examined and, if any feature has an MSA of ≤ 0.5, the feature with the lowest MSA is removed. The KMO index is then re-calculated and this process of omitting the variable with the lowest MSA value is continued until all features reach an MSA value of ≤ 0.5.

An additional step that is often taken to ensure that the results of factor analysis methods are robust consists in removing variables with low final communalities from the analysis. As Lee (2000: 244ff.) points out, this is not something that Biber (1988) did (though Biber 1995 shows awareness of the issue)[54]. Communality is measured as the sum of all the squared factor/component loadings for any one variable and therefore refers to the proportion of variance within a variable that is explained by the extracted factors. In other words, a low communality indicates that a substantial proportion of a variable's variance is not accounted for by the reduced solution. There are no hard and fast rules as to what constitutes a reasonable communality cut-off point because it very much depends on how much total variance a solution explains; however, in the context of MDA, Biber (1995: 138) recommends eliminating linguistic features with communalities ≤ 0.20. This is also the cut-off point that is used in the present chapter. As Lee (2000: 246–247) points out, many features with low final communalities are also low-frequency features.

Once these various steps have been undertaken to eliminate very unevenly distributed features, those with overly high or particularly low correlations, and low communalities, the overall MSA can be re-calculated to evaluate the suitability of the

[54] According to Lee (2000: 244ff), eight of Biber's (1988) features had final communalities of ≤ 0.20 yet were still retained in the final model.

dataset for this kind of analysis. The resulting overall KMO values may be interpreted following Kaiser & Rice's (1974: 112) wonderfully flamboyant approximate scale:

$\geq .90$ marvellous
$\geq .80$ meritorious
$\geq .70$ middling
$\geq .60$ mediocre
$\geq .50$ miserable
$< .50$ unacceptable

Arguably more meaningfully, KMO values may be compared to those of previous MDA studies. Unfortunately, however, few MDA studies report these: out of the 230 MDA studies that Goulart & Wood (2021: 124) surveyed, 26 claim to have checked the factorability of their data using KMO and 24 report exact overall KMO values. For these, Goulart & Wood (2021: 124) calculate a mean KMO value of 0.69 ($SD = 0.08$, $\min = 0.43$, $\max = 0.86$), which would suggest that correlation matrices typically entered in MDAs are only "mediocrely" to "middlingly" suitable for factor analysis. However, all of the suggested methodological advancements on the traditional MDA framework outlined so far, in particular the removal of "obvious" correlations, the elimination of highly unevenly distributed features, those that have low MSA scores or low communalities, and the transformation of particularly skewed distributions ought to contribute to higher overall KMO values and to correlation matrices that are more suitable for this kind of data-reduction analysis.

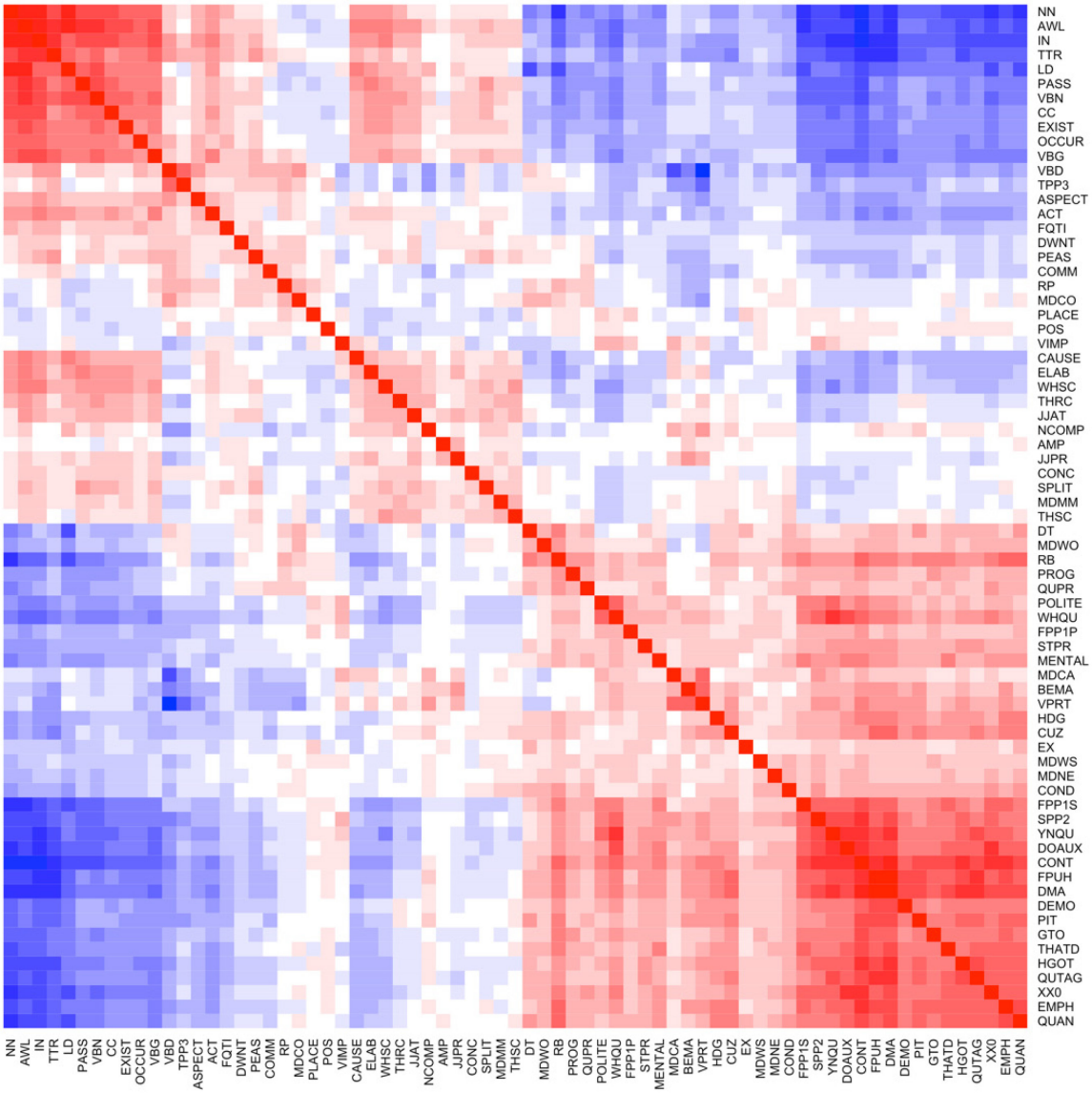


Fig. 53 Correlation matrix of the signed log standardised relative frequencies of the features analysed in the second PCA-based MDA (see 7.3.2)

7.2.7 Arbitrary thresholds in the computation of dimension scores

The MDA framework also foresees the computation of dimension scores (sometimes also referred to as 'factor scores') for each text in the corpus under study. Mean dimension scores for specific (sub-)registers or other subgroups of the corpus can then be compared. As explained in 6.2.4, Biber (1988: 93) and most other linguists following his method exclude features with factor loadings below a pre-determined cut-off point from the computation of their dimension scores (as in Biber's original MDA study, most use ±0.35). In addition, if a feature loads onto more than one factor with a loading above the chosen cut-off point, the feature is only counted for the factor on which it has the highest loading. To calculate the dimension scores that correspond to a particular factor, the standardised normalised frequencies for each of the salient positive-loading features on that factor are added together whilst the salient negative-loading ones are subtracted.

Biber's MDA framework has been criticised for relying on an arbitrary cut-off point to include or exclude certain features from its dimension scores (e.g., Evert 2018: 12). There is, however, a valid rationale for removing low-loading features from dimension scores: they are likely to simply reflect noise in the data that is arguably best removed from models that aim to be representative of a larger population. Statistical methods can be applied to exclude non-significant loadings, i.e., those that are likely to be the result of random patterns of variation in the data (Husson et al. 2018: 220). However, with large data sets such as those typically used in MDA studies, such significance tests are likely to return extremely low thresholds. In relying on a cut-off point of ±0.35, Biber (1988: 93) applied a slightly more conversative version of the common threshold in social sciences of ±0.30. It has the advantage of excluding loadings that, whilst perhaps statistically significant, may not have any practical relevance as they account for less than 12.25% ($= 0.35^2$) of the shared variance (Lee 2000: 207).

Biber's exclusion of features contributing to more than one dimension has also been criticised. On the one hand, it has the advantage of making the dimension scores "experimentally independent, as each feature contributes to only one dimension" (Lee 2000: 209; see also Biber 1988: 93). However, it clearly adds a degree of arbitrariness: for instance, in Biber's (1988) model, past participle WHIZ deletions contribute to Dimension 5 scores, but not to Dimension 1 scores, even though in Biber's rotated factor solution, past participle WHIZ deletions contributed to a very similar extent to Factor 1 (-0.39) and Factor 5 (0.43). Moreover, Biber (1988: 85) advocated for the use of oblique rotation because linguistic dimensions can reasonably be expected to intercorrelate given that "from a theoretical perspective, all aspects of language use appear to be interrelated to at least some extent" (Biber 1988: 85 fn.). If this is the case, it should come as no surprise that individual linguistic features may also make significant contributions to more than one dimension.

Finally, the method that Biber and others following the MDA framework have traditionally used to calculate dimension scores essentially involves dichotomising all feature loadings: all features that have loadings above the chosen cut-off point (while not contributing more to another factor) contribute equally to the dimension scores, whilst those that are excluded from the dimension scores do not contribute at all. The feature contributions are therefore equal to either one or zero. This means that all loadings above the cut-off point are considered equally important, even though they may actually have made substantially different contributions to the original factor solution. For example, in Biber's (1988) model, the standardised frequencies of past tense verbs and present participial clauses are treated as equally important contributors to Dimension 2 scores, even though their factor loadings are widely different (0.90 vs. 0.39). Hence, this approach can be argued to grant less salient linguistic features disproportionate significance. At the same time, however, such a dichotomous approach to calculating dimension scores is not without its advantages.

In non-linguistic uses of factor analysis, discarding the relative importance of features has been shown to distort results only marginally. In fact, dichotomisation increases the chances that the resulting dimension scores may be replicated with a new sample of texts because it essentially removes some of the random noise inherent to small differences between factor loadings (Gorsuch 2014: 275–276).

As mentioned in 7.2.5, with PCA, the component feature loadings are correlation coefficients between the observed features and the components and are therefore much simpler to interpret than the factor loadings computed in factor analyses. Nonetheless, the question still arises as to how much relevant information the exact feature loadings contribute to any component, whether a cut-off point is needed and, if so, which one. In effect, three solutions to calculate dimension scores can be envisaged. The first solution simply consists in using the loadings as they are. In other words, on any one component, the standardised normalised feature frequencies of any one text are multiplied by their respective loadings on this component and these values are added to compute dimension scores. The second consists in applying a cut-off point to exclude low-loading features whilst retaining the other loadings as multiplying factors to calculate the dimension scores. Finally, the third solution is the one typically adopted in MDA studies: as explained above, it consists in dichotomising loadings according to a cut-off point. With a cut-off point of ±0.30, this would mean that to calculate dimension scores only the unweighted standardised normalised features with loadings of $\geq$ 0.30 or $\leq$ -0.30 are added (they are in effect multiplied by one, whilst those with loadings between -0.30 and 0.30 are multiplied by zero).

If the aim of an MDA is to produce a model of linguistic variation that is generalisable beyond the sample under study, as was presumably the case with Biber (1988), then solution three may well be the wisest. However, the potential issues it causes downstream should not be downplayed. For a start, the resulting dimension scores no longer correlate perfectly with the factors/components they purport to quantify. This is why Lee (2000: 211) argues that they should really be referred to as "estimates" rather than "scores" (see also Child 1990). Moreover, the reported R^2 are no longer true. Hence, whilst Biber's (1988) Factor 1 accounts for 26.8% of the shared variance (Biber 1988: 82–83), this is not true of Dimension 1 that explains considerably less due to the loss of information caused by the dropping of low-loading features, the use of dichotomous loading weights (1 or 0) for the remaining features, and the exclusion of features with significant loadings on more than one factor. Additionally, whilst we have seen that having a cut-off point is not necessarily a bad idea, it nevertheless involves an arbitrary decision that will inevitably exert substantial influence on the resulting dimension scores and can therefore potentially be a potential source of bias (see also Evert 2018). In light of these issues, the first solution was chosen. For the reasons explained above, the MDAs presented in this chapter also allow linguistic features to contribute to more than one dimension. Whilst this may somewhat

complicate the interpretability of the resulting dimensions scores, it is in line with our understanding of communicative processes in which linguistics features are expected to intercorrelate in many ways. The chosen solution bears the advantage of not relying on an arbitrary cut-off point, maintains the true correlations of features to dimensions and thus does not distort the PCA solution. Since the loadings themselves act as factors in the computation of the dimension scores, the amount of noise added by low-loading features is assumed to be negligible.

7.2.8 Misleading visualisation of results

To date, the results of MDAs have almost exclusively been visualised by plotting the mean dimension score of (sub-)registers on a vertical line representing a dimension cline with the most negative-scoring text categories placed at the bottom of the line and the highest scoring ones at the top.[55] This is how Biber first visualised the results in his seminal 1988 MDA study (see Fig. 29 for a reproduction of such a plot). This section explains why such visualisations are not satisfactory.

First, plotting only mean dimension scores entirely ignores intra-category variability. It gives no indication of the range of dimension scores covered by a single (sub-)register category and tacitly suggests that the distributions of dimension scores within each category are all normal. In fairness, most MDA studies do report standard deviations alongside these mean dimension scores – however, usually only in tabular form, which makes it is very difficult to grasp how much overlap there is between different (sub-)register categories. Whilst it is perfectly understandable that digital plotting methods were more limited in the late 1980s, such dimension plots are still widely used in MDA publications today. For example, Gardner et al. (2019) present the results of an MDA study on learner academic writing with plots such as Fig. 54. An improvement on Biber's (1988) dimension plots is that, on Fig. 54, the number of texts that constitutes each category is printed in brackets; however, the plot nonetheless provides no indication of the distribution shape of dimension scores in each category. Furthermore, it is impossible to gauge whether any of the observed differences between the categories are likely to be statistically significant. This is why, in Chapter 6, the dimension scores of each individual text in the corpora were plotted, thus helping the reader to easily evaluate and compare the number of texts in each category, as well as to detect outliers. These 'raincloud plots' (see, e.g., Fig. 31) also include boxplots for ease of comparison between the different categories: these display the category median value (which is less susceptible to outliers than the mean) and the interquartile range, which, in combination with the individual text points, gives a good impression of the dispersion of the texts along any one dimension. Given "the centrality of text in corpus-linguistic inquiries" (Biber 2021: n.p.) and the fact that

[55] Some studies (e.g., Lee 2000) have visualised the results of MDA using barplots of mean dimension scores, which is arguably even more misleading.

the MDA framework applies a text-linguistic research design (Biber et al. 2016: 357), being able to visualise the position of each text on any one dimension is crucial.

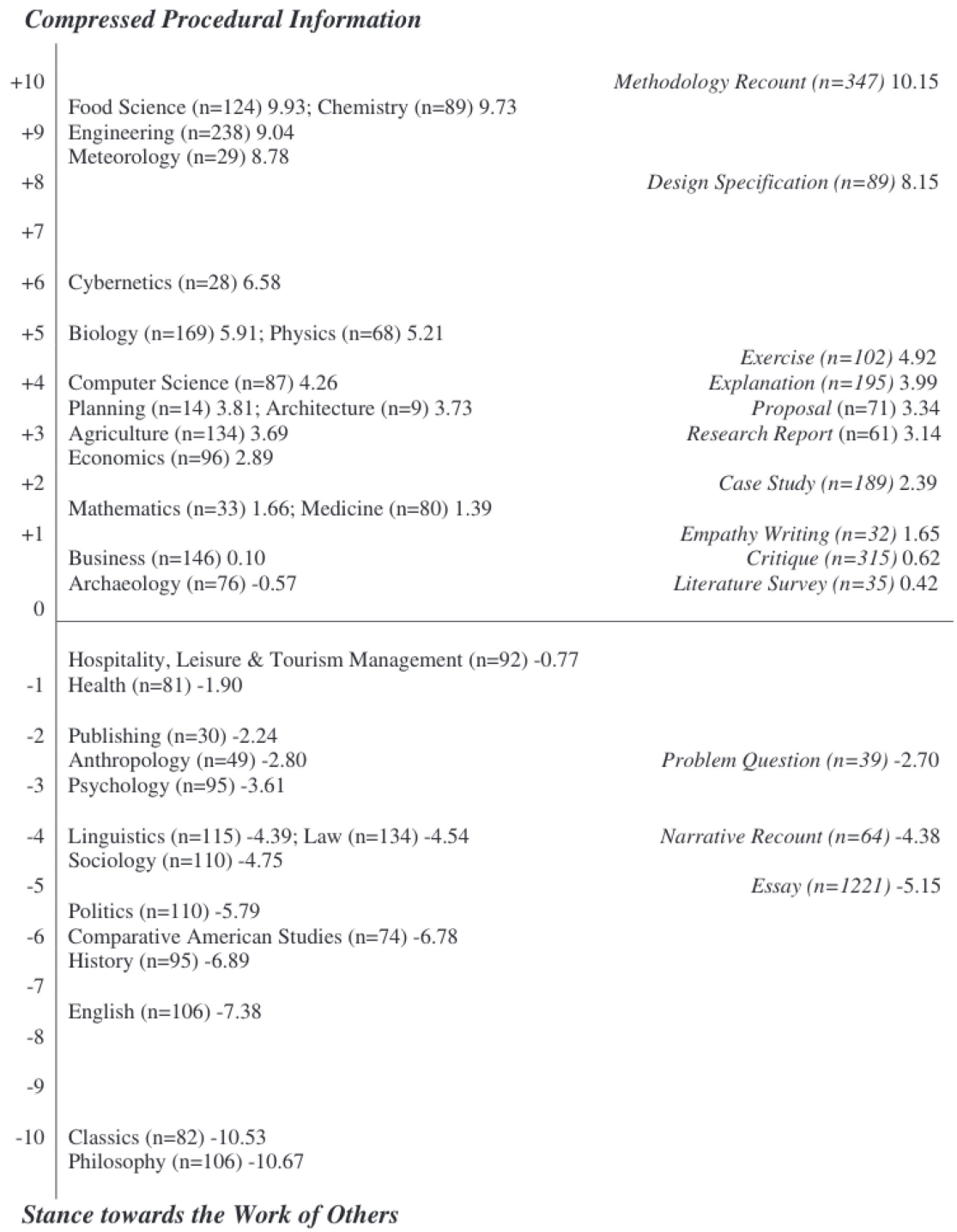


Fig. 54: Dimension 1 mean scores for disciplines (left) and genre families (right) from Gardner et al.

The second issue, common to plots such as Fig. 54 and the raincloud plots of Chapter 6, is that they encourage the reader to only consider linguistic variation on a single dimension at a time. In the case of an additive MDA that compares the dimension scores of previously confirmed individual dimensions, as in Chapter 6, this

is not necessarily a problem. If, however, the results of MDAs are to be genuinely understood and interpreted as multi-dimensional, it is crucial that the position of texts be compared along all the potentially relevant dimensions. In practice, of course, humans struggle to visualise more than two or three dimensions at once. That said, facetted plots such as Fig. 57 can be used to visualise clusters of texts in all the possible combinations of dimensions. On these, however, small differences between text categories are very difficult to discern. This is why this chapter makes extensive use of 2- and 3-D plots that display the position of individual texts on two or three dimensions at once.

Similarly, most MDA studies report lists of salient features and their loadings on individual dimensions in tabular format (e.g., Table 40). Not only do these tables usually exclude the features with loadings below the arbitrarily chosen cut-off point (see 7.2.7), they also do not readily allow for any comparisons of loadings across different dimensions. By contrast, the tables of loadings in this chapter include all the features entered in the analyses. Cell shading and font colours are used to highlight the most important contributions (see Tables 68 and 69). Furthermore, graphs of features (e.g., Fig. 59) are used to visualise the linguistic features that contribute most to the position of texts on biplots (more on how to interpret these in 7.3.1). Again, these graphs allow for a better visualisation of the extent to which some features make salient contributions to more than one dimension.

Thus, in terms of visualisation, too, the method proposed in this chapter bears many similarities with Neumann & Evert's (2021) geometric multivariate analysis (GMA) approach which is, in turn, inspired by Biber's MDA framework, "but takes a more geometric perspective emphasizing the visualization of individual texts in multidimensional feature space" (Neumann & Evert 2021: 148). Linguistic differences between texts are analysed by examining the Euclidean distances between texts in this multi-dimensional feature space.

7.2.9 Difficulties in establishing the statistical significance and robustness of the results

Evert (2018) suggests using bootstrapping and/or cross-validation as a means of assessing the significance of the results emerging from PCAs. Following a similar logic, Lee (2000: 346–369) ran EFAs on subgroups of his data to check the robustness of the results. In the present chapter, the PCAs will be re-run on random subsets of two-thirds of the data for the same reason.

Issues pertaining to the way in which the significance of the register differences between dimension scores emerging from MDA studies are usually tested have already been addressed in 6.2.6. It explained why, in both Chapter 6 and the present chapter, linear mixed-effects models were computed to compare the dimension scores of the

results of the MDAs to account for multiple variables that may contribute to linguistic variation (register and textbook proficiency level), the fact that interactions of these variables may also make significant contributions, and for the non-independence of texts sourced from the same textbook series, novel, or website (see Candarli 2021 for a similar approach with nested data in a learner corpus). The approach described in 6.2.6 is also followed in the present chapter.

7.2.10 Issues related to reproducibility and replicability

Though the terms are sometimes used interchangeably and different (at times incompatible) definitions abound, in computational sciences, 'reproducibility' usually refers to the ability to obtain the same results as an original study using the authors' data and code, whilst 'replicability' refers to obtaining compatible results with the same method but different data (Association for Computing Machinery 2020; see Berez-Kroeker et al. 2018 for similar definitions in the context of linguistic studies).[56] As Popper wrote as early as 1935, "non-reproducible single occurrences are of no significance to science" (English edition from 1959: 86). In other words, sound science requires the results of studies to be reproducible (providing that the data can ethically and legally be made available) or at least conceptually replicable. In other words, if:

> a scientist publishes the results of an experiment, there should be enough of the methodology published with the results that a similarly-equipped, independent, and skeptical scientist could reproduce the results of the experiment in their own lab (Gezelter 2009: n.p.).

In the conceptual sense of replication, issues pertaining to the robustness of the results (see 7.2.9) are directly connected to those related to the replicability of the analyses. Of course, it goes without saying that, ideally, direct replication of MDA studies (reproducibility) also ought to be possible. For the vast majority of MDAs, even those analysing publicly available corpora such as Brown, LOB or the BNC1994, this is currently not the case because the most popular software for tagging and counting linguistic features for English MDA studies, the Biber Tagger (Biber 1988; 2019), is not (at the time of writing[57]) available to the wider research community, either under an open-source licence or commercially. Though the explanations of its algorithms in Appendix II of Biber's 1988 monograph have rightly been praised for their comprehensiveness, when Nini (2014; 2019) attempted to replicate the Biber Tagger, he came across a number of unclear cases. Though the MAT (see 6.2.3) has been shown to produce results highly comparable to the Biber Tagger in its 1988 version, unexplained differences nonetheless remain. Moreover, Biber has continued to improve the tagger with "several major rounds of revision and extension, resulting in

[56] Note that other terms are also frequently used to refer to the same or related concepts, e.g., *repeatability*, *robustness* and *generalisability* (cf. Belz et al. 2021: 2–3).

[57] Some of Douglas Biber's collaborators have been working on a new version of the Biber Tagger (to be named differently) with the aim of making available via a "semi-public" online interface (Douglas Biber, personal communication 2018).

the analysis of a much more comprehensive set of linguistic features" (Biber 2019: 14) for which no details of the algorithms are publicly available. Thus, to ensure that the results of an MDA are reproducible, it is crucial that the computer programme(s)/the algorithms used to tag and count the linguistic features entered in the MDA are published or made available on request. This is also one of the reasons why McEnery & Hardie (2011: 112) conclude that replicability remains "something of a concern for the MD [multi-dimensional analysis] framework".

The Multi-Feature Tagger of English (MFTE) used in the analyses of the present chapter has been released under a GNU General Public licence (see Online Appendix 7.1) and can therefore be scrutinised by the wider research community. Whilst the corpora cannot be released for copyright reasons, the full tabular results of counts as output by the MFTE are included in the Online Appendix 7.2–7.7, as well as the full code to reproduce the PCA results, the statistical tests, and all the figures on the basis of this count data. This is important because, for all the complexity of advanced statistical methods such as those typically used in MDAs, it is worth remembering that such methods nevertheless require researchers to make countless (as we have seen, often rather arbitrary and certainly always subjective) decisions on a host of parameters. Given that "there is nothing sacred about one particular factor [or component] solution or one grouping of features" (Lee 2000: 370), it is crucial that each parameter choice be transparent, and that the robustness of the results can be checked by independent researchers. In addition, and as recommended by Lee (2000: 393; see also Biber 1990), the present results were replicated on various subsets of the data to test the replicability and robustness of the models presented and interpreted in the present chapter (see 7.4).

7.3 Results

As in Chapter 6, register variation within the TEC is first explored in 7.3.1 before the three main textbook registers, Conversation, Fiction, and Informative texts, are mapped against the three corresponding reference corpora in 7.3.2.

7.3.1 Linguistic variation within the TEC

This section focuses on register variation within secondary school EFL textbooks. The texts of the TEC (as defined in 7.2.1) were all tagged with the MFTE (see 7.2.2–7.2.5). Of the tagger's three outputs, the complex normalisation table was used as the basis for the PCAs of this section (see 7.2.4). Three features were removed: CD (cardinal numbers) because numbers had to be removed from most textbook texts due to the presence of line numbering and footnote numbers for glosses, as well as the variables LIKE and SO because these categories were created to increase the precision and recall of other linguistic features and cannot be meaningfully functionally interpreted (Le Foll 2021e: 17–18, 33–34). Next, applying the feature

exclusion procedure described in 7.2.6–7.2.7, the counts for the BE *(un)able to* construction (ABLE) were merged with the category of predicative adjectives (JJPR), whilst the counts for passive GET constructions (PGET) were added to the BE passive counts to create a more general passive category (PASS) because both categories were absent from more than two-thirds of texts. Other features which were also only observed in a third or fewer of TEC texts but could not be meaningfully subsumed with any other features were excluded from this MDA. These are: CONC, DWNT, ELAB, EMO, GTO, HGOT, HST, MDMM, PRP, QUTAG and URL (see Appendix I for the full table of features and the code in the Online Appendix 7.4 for details of the procedure). The iterative process to arrive at individual feature MSA values of > 0.5 (described in 7.2.6) led to the exclusion of one additional feature: MDWS. Finally, four features were removed on the basis of their low final communalities: STPR, MDNE, HDG and CAUSE. In addition, sixteen outlier texts were removed on the basis of some extremely high feature counts (see code in Online Appendix 7.4 for details).

The following PCA is therefore based on a matrix of 1,961 texts by 61 features, all z- and signed log-transformed (see 7.2.6), with a satisfactorily high overall KMO factor adequacy index of 0.88, or "meritorious" according to Kaiser & Rice (1974: 112). To determine the number of components to be considered in the analysis, a scree plot was first generated, see Fig. 55. It shows the amount of variation each component captures from the TEC data. The "elbow" method (Jolliffe 2002: 115–118) is difficult to apply here because the plot can be said to feature several "breaking points". Following Biber's (1988: 84) advice to extract more rather than fewer components to start off with, the first six components were originally retained for further analysis. Together, these account for 50.88% of the total variance.[58]

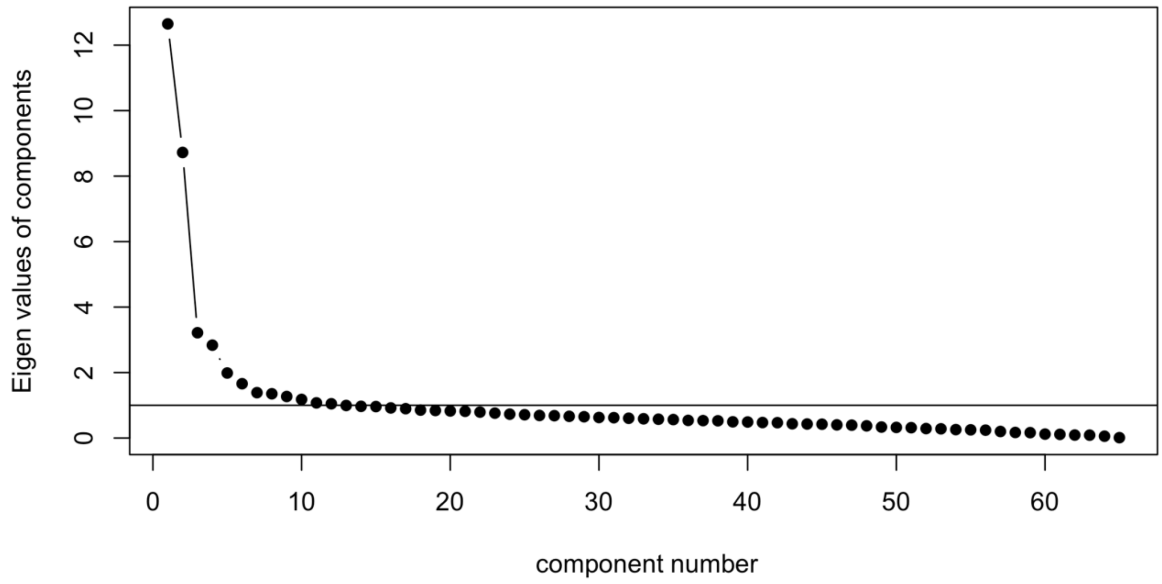


Fig. 55: Scree plot of the eigenvalues of the PCs for the TEC data

[58] For reference, Biber's (1988) EFA solution of seven factors accounted for 51.9% of the shared variance (Biber 1988: 83).

The distribution of the texts on the first three components was first explored interactively in a 3-D visualisation computed using the {pca3d} R package (see Online Appendix 7.6 and snapshots in Fig. 56). Here, and on all subsequent scatterplots, every data point represents a single text from the TEC. The closer points are, the more linguistic similarities they share. At first sight, the most striking aspect of the 3-D visualisation is that there are two clearly separated clusters of texts: one consisting of instructional language (in yellow) and the other of the remaining textbook registers. Within this second, much larger cluster of texts, we find that conversation is concentrated at one end (in red) and informative texts (blue) at the other, with fiction (green) and personal correspondence (purple) interspaced in between. The 3-D visualisation makes clear that all three components contribute to distinguishing register-based intra-textbook variation.

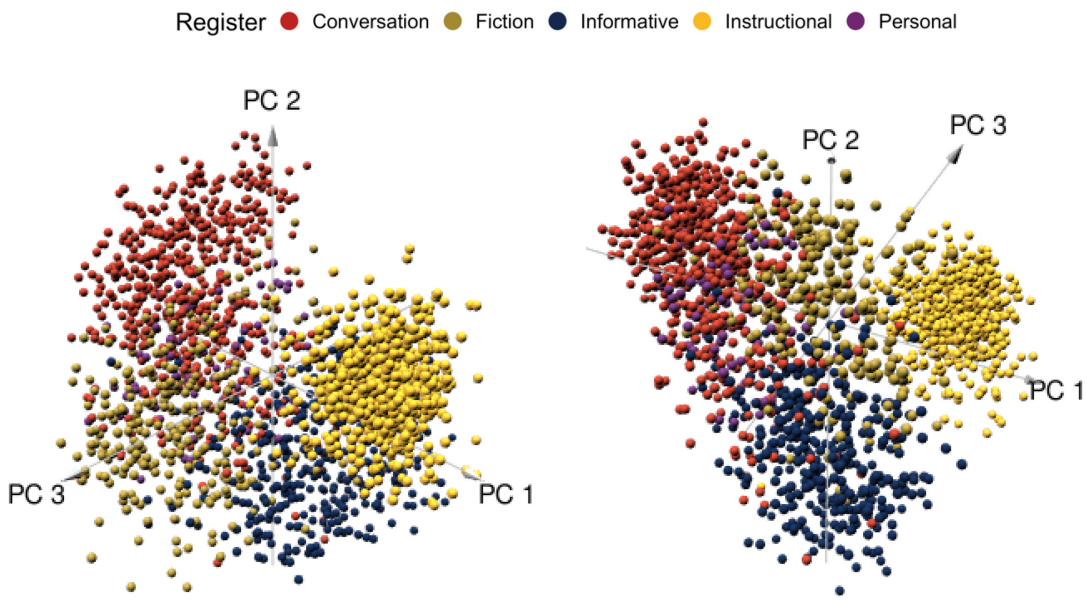


Fig. 56: Snapshots from the 3-D visualisation of the first three dimensions of the multi-dimensional model of intra-textbook variation

By contrast, the remaining three dimensions (PC4, PC5 and PC6) do not appear to distinguish between different textbook registers. This is illustrated in the biplot matrix of all combinations of the six retained dimensions in Fig. 57. The same colour scheme is used to encode the different registers as in the 3-D visualisations and, in addition, the proficiency levels of the textbooks from which each text stems are represented by different shapes: beginner textbook texts (level A) are assigned the circle shape, while the texts from the most advanced textbooks in the TEC (level E) are represented by diamonds (see Online Appendix 7.2 for zoomable version of Fig. 57).

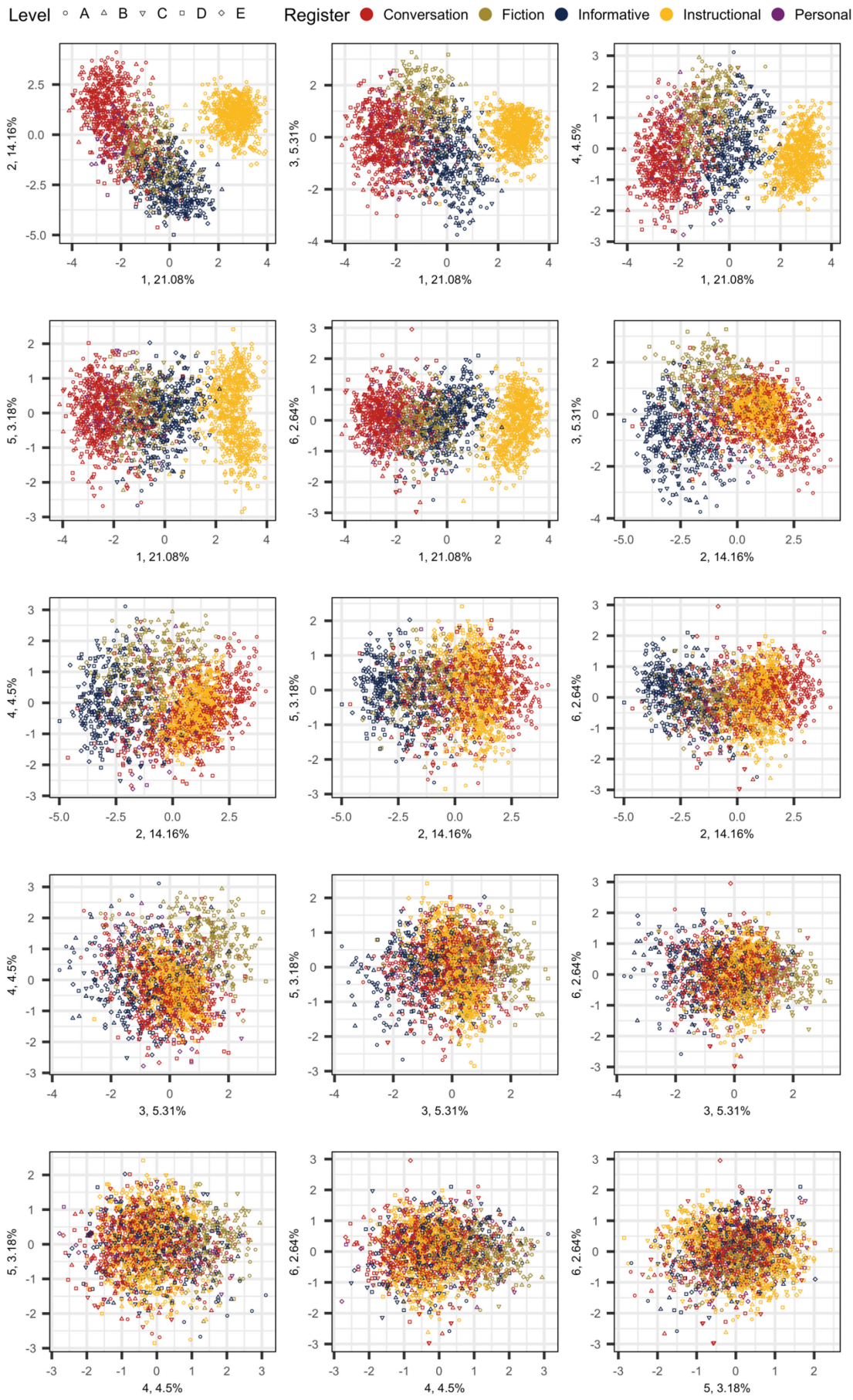


Fig. 57: Scatterplot matrix of combinations of the first six dimensions of the model of intra-textbook variation (the number before the comma on each axis label shows

which principal component is plotted on that axis; this is followed by the percentage of the total variance explained by that particular component)

Fig. 58 is a more fine-grained projection of the texts of the TEC on the first two dimensions which, together, account for the greatest linguistic differences between the different registers of the TEC. The ellipses represent the 95% confidence intervals around each of the five textbook register centroids. As already observed in the 3-D plot, two clusters of texts are evident: to the right of the plot, instructional texts form a tight cluster whose ellipse does not overlap with any of the other four textbook registers. Thus, we can conclude that instructional language has a very characteristic linguistic profile which clearly sets this register apart. The linguistic features that contribute most to this very specific profile can be seen in the top right panel of Fig. 59, in the area of the plot that corresponds to the area where the cluster of instructional texts can be found in Fig. 59. They are, as illustrated in (215), imperatives (VIMP), verbs of communication (COMM), and verbs depicting mental processes (MENTAL).

(215) **Look** at the other groups' guides and **choose** which channel you would like to watch. **Use** the key phrases for making and **justifying** a choice. **Work** in pairs. **Answer** the questions. <TEC: Solutions pre-intermediate>

In addition, second person referents (SPP2) and WH-questions (WHQU) also feature in the upper-right panel of Fig. 59 and are thus very typical of instructional language, too, e.g., (216); however, these features are situated closer to the *y*-axis as they also make strong contributions to the positive end of the model's second dimension (PC2), which, much like Biber's (1988) Dimension 1 (see Fig. 29), corresponds to an involved vs. informational language continuum.

(216) Reactions
a) Describe **what happens** in the second half of the story (after line 43). How **do** the customers **react**? How **does** the narrator **react**?
b) **Do** you **understand** the way they **react**?
Short stories often **start** unusually ("medias in res" - right in the middle of the action) **and end** with a surprise. Look at "Deportation at breakfast" again: find these elements **and** say **why** they **are** important here. <TEC: Green Line 5>

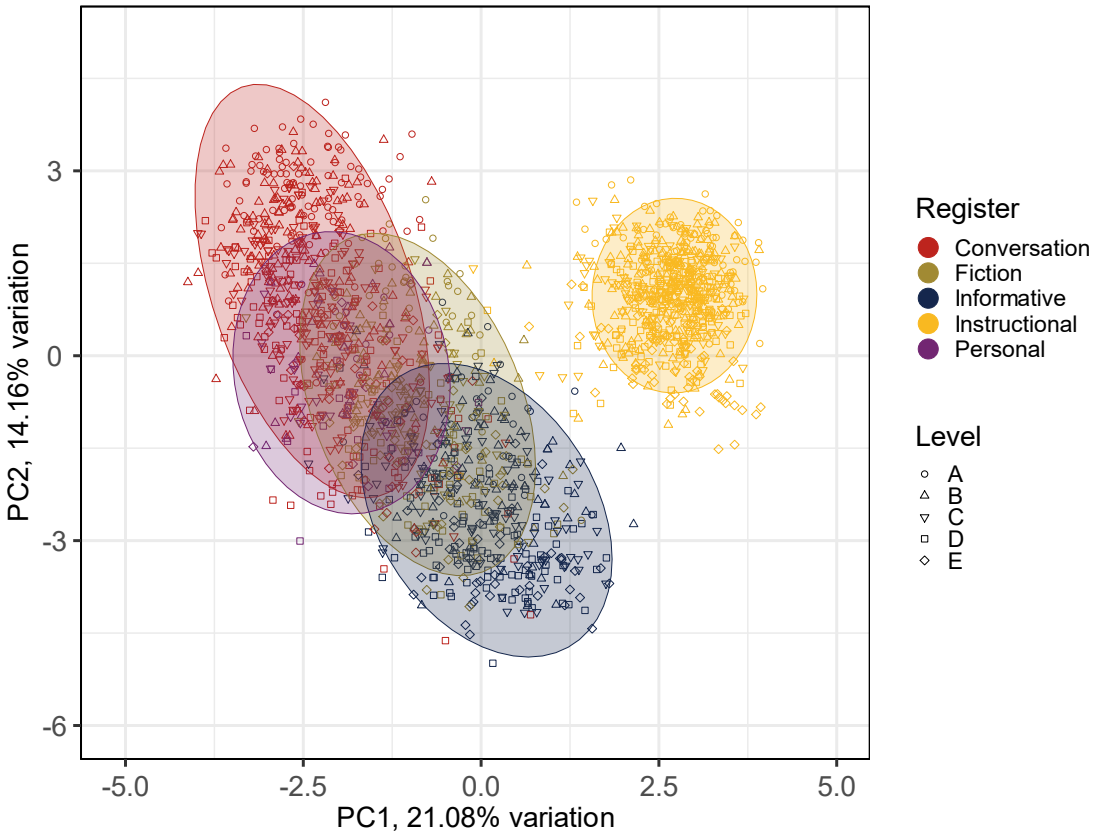


Fig. 58: Projection of the texts of the TEC on the first and second dimensions of the model of intra-textbook variation

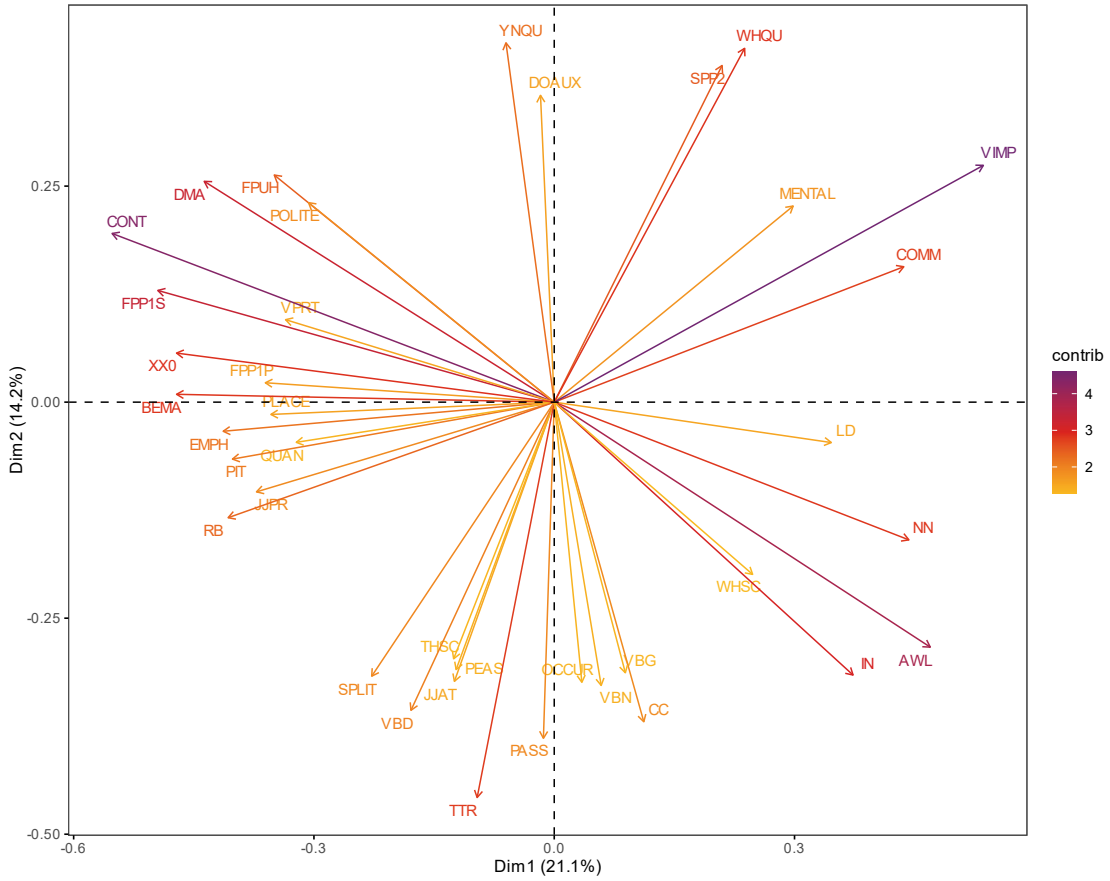


Fig. 59: Graph of the features with the strongest contributions to the first and second dimensions of the model of intra-textbook variation

The second, larger cluster on Fig. 58 reveals a clear register cline with textbook texts depicting conversations towards the top-left end of the cluster, fiction, and personal correspondence in the middle, whilst informative texts are concentrated at the bottom-right end of the cluster. The features that contribute most to this distribution of registers can be seen in the graph of features on Fig. 59. The advantage of this kind of biplot is that we can see that the features that are on or very close to one of the two axes contribute principally to just one of these first two dimensions, whilst those that draw diagonals contribute to both dimensions. Thus, Fig. 59 shows that the upper end of the large, "non-instructional textbook English" cluster is characterised by high frequencies of fillers and interjections (FPUH), markers of politeness (POLITE), discourse markers (DMA), verbal contractions (CONT) and present tense verbs (VPRT), e.g., (217). By contrast, the lower end of the cluster features texts with longer words (AWL), high frequencies of nouns (NN), prepositions (IN) and subordinate WH-clauses (WHSC), e.g., (218). These two extremes echo the features with the highest estimated factor loadings on the two ends of Biber's (1988) 'Involved vs. Informational Production' dimension (see Table 40).

(217) Can I help **you?**
Yes, **have you** got the new 'Pets' magazine, **please**? I can**'t** find it.
It's there - next to the sports magazines.
Excuse me. **Where can** I try on this sweatshirt**?**
There, on the left.
Thanks. I **like** the colour, but the size **isn't** right.
No problem. We**'ve** got other sizes, too. <TEC: Green Line 1>

(218) The **Aboriginal Memorial** is an **installation of** 200 hollow **log coffins from Central Arnhem Land**. **Artists** made it to commemorate all the indigenous **people** who, since 1788, have lost their **lives** defending their **land**. **Visitors** can see it **in** the **National Gallery of Australia**. The **artists** said the **museum authorities** must locate this **installation in** a public **place** where they could preserve it **for** future **generations**.
<TEC: New Missions 2^{de}>

The large cluster's slanted shape on Fig. 58 indicates that both the first (PC1) and second (PC2) dimension of this multi-dimensional model of intra-textbook variation capture important aspects of register-based or situational variation. The biplot of the first two dimensions, Fig. 58, clearly shows that many linguistic features (foremost those plotted on Fig. 59) are distributed quite differently across at least three out of the five textbook registers under study: this is illustrated on Fig. 58 by the fact that instructional language forms its own very distinct cluster, and the ellipses of the conversational and informative texts overlap very little. The ellipses for the fiction and personal correspondence texts, however, overlap much more, suggesting that these two textbook registers are not readily distinguishable on these first two dimensions (though see 7.3.1.1 and 7.3.1.2), which brings us to the third and fourth dimensions.

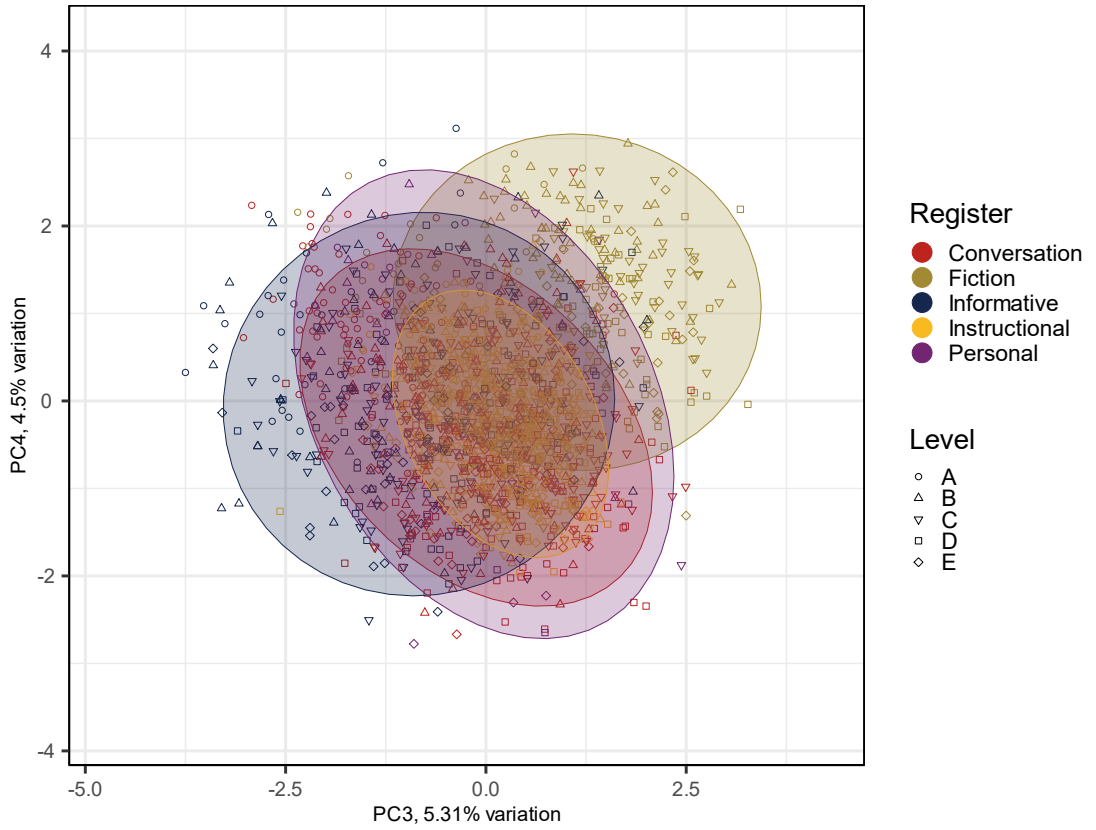


Fig. 60: Projection of the texts of the TEC on the third and fourth dimensions of the model of intra-textbook variation

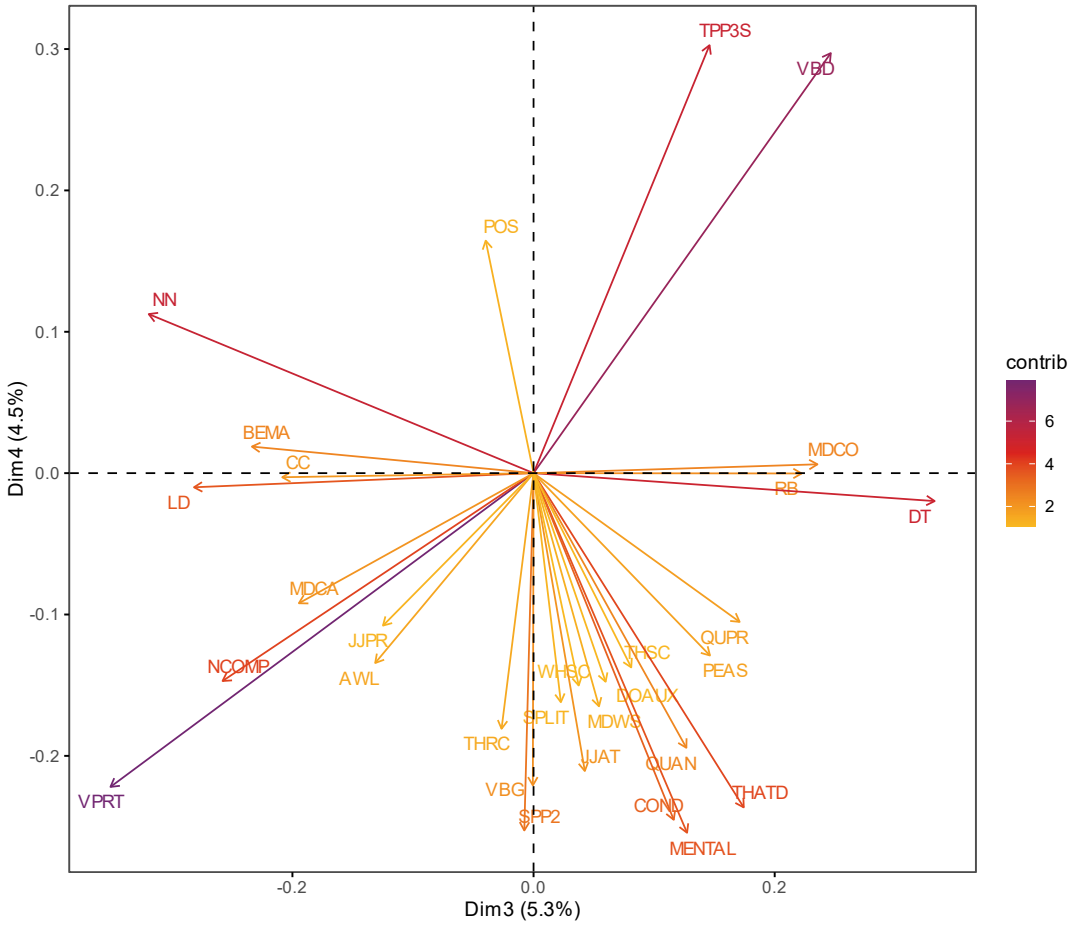


Fig. 61: Graph of the features with the strongest contributions to the third and fourth dimensions of the model of intra-textbook variation

Fig. 60 displays the positions of the texts of the TEC on the third (PC3) and fourth dimensions (PC4). The intersection of these two dimensions highlights a distinctive linguistic profile for at least some of the fiction texts of the TEC. Indeed, part of the green ellipse is set apart from the rest of the texts. The features that contribute most to this characteristic linguistic profile can be found in the top right panel of corresponding graph of features on Fig. 61. Just like on the narrative end of Biber's Dimension 2, the frequency of past tense verbs (VBD) and third-person references (TPP3S) make the largest contributions to this characteristically "narrative" cluster.

A closer look at the shapes of the points in this non-overlapping portion of the textbook fiction ellipse in the top-right panel of Fig. 60 reveals that it is foremost composed of narrative texts from intermediate to advanced textbooks (levels C to E). To explore this further, Fig. 62 displays the texts of the TEC on the same two dimensions as in Fig. 60 but this time the colour scheme and the ellipses correspond to the proficiency levels of the textbooks from which the texts have been extracted, rather than the register of each text (which is, instead, coded by the shapes of the points). A comparison of the two biplots (Fig. 60 and Fig. 62) shows that, whilst register-based variation is greater, textbook proficiency level also makes some notable contributions to linguistic variation in Textbook English, as evident on the third and fourth dimensions. These different factors contributing to linguistic variation in the TEC will be tested in more in-depth analyses of the first four dimensions in 7.3.1.1–7.3.1.4.

The fifth and sixth dimensions (PC5 and PC6), however, will not be examined any further as they account for comparatively little of the total variance (PC5 = 3.18% and PC6 = 2.64%). Both the visualisations (Fig. 63 and the 3-D projections of PC4–PC6 in the Online Appendix 7.6) and the mixed-effects models conducted to explore these dimensions (see Online Appendix 7.6) indicate that they contribute very little to differentiating between different text registers, proficiency levels, or textbook series.

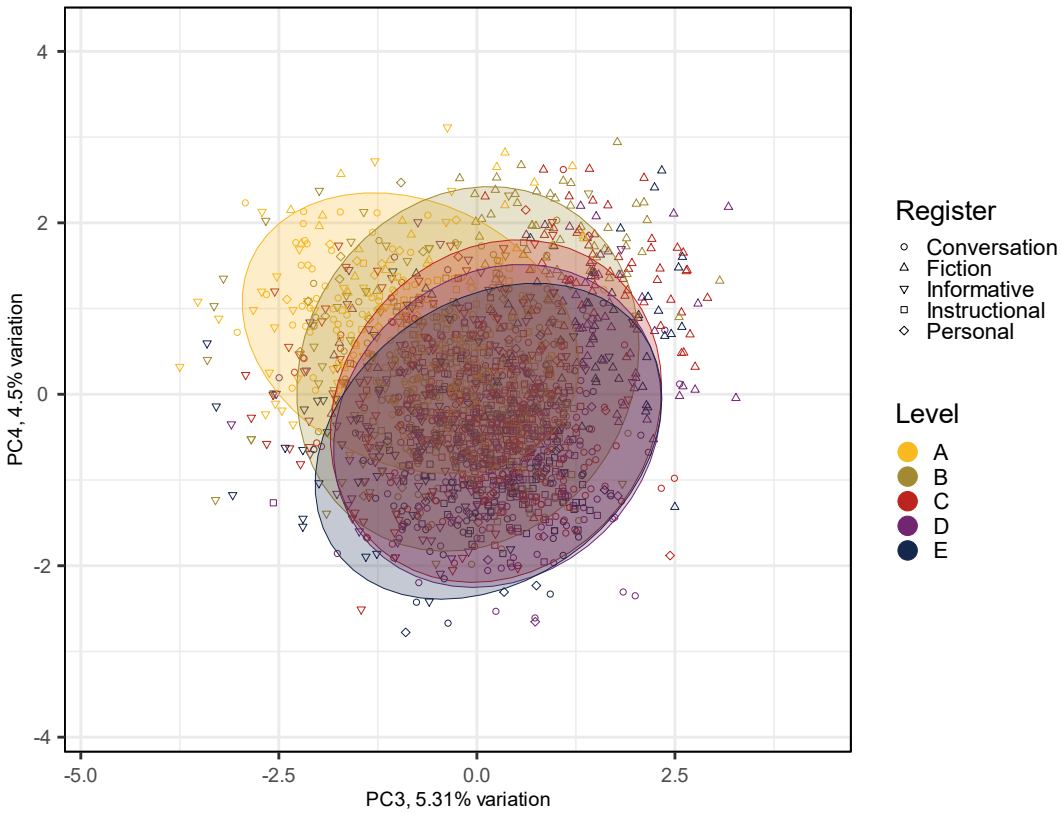


Fig. 62: Projection of the texts of TEC on third and fourth dimensions with colours and ellipses indicating the proficiency level of the textbooks (as opposed to register as in Fig. 60)

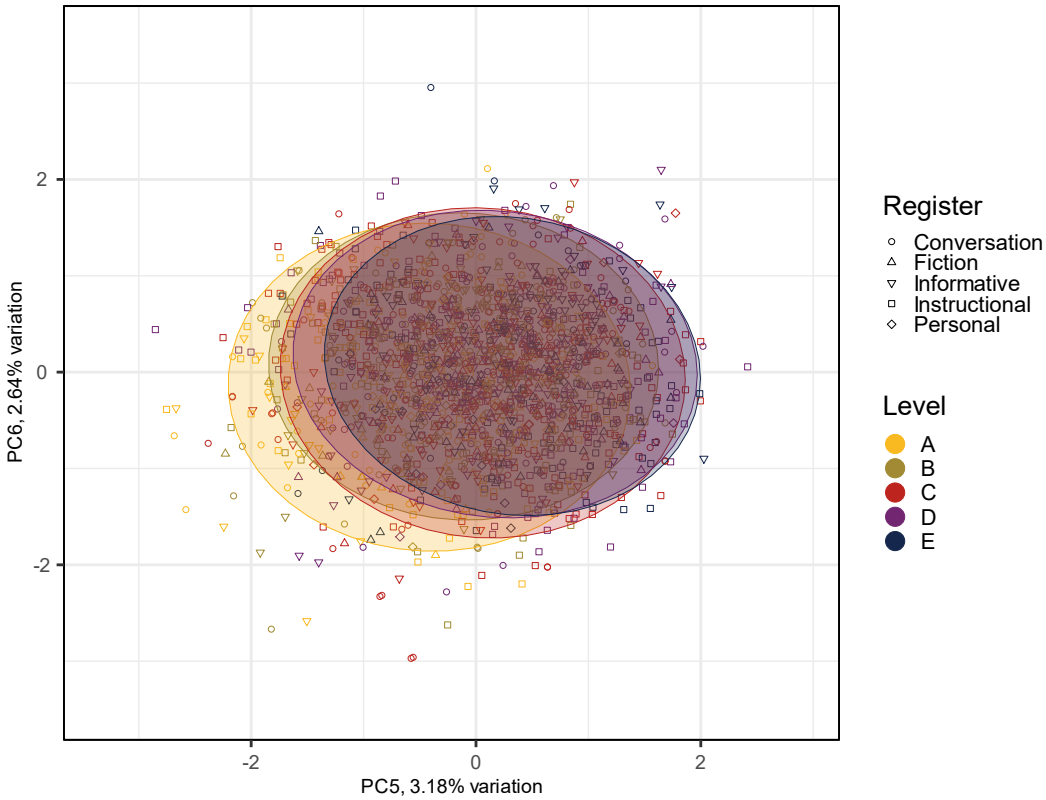


Fig. 63: Projection of the texts of the TEC on the fifth and sixth dimensions of the model of intra-textbook variation

7.3.1.1 Variation along Dimension 1: 'Overt instructions and explanations'

As we saw in Fig. 56–58, the first dimension to emerge from this PCA-based model of intra-textbook variation primarily separates instructional texts and explanations from the rest of the TEC data. Given that the negative end of the dimension does not imply any specific text quality other than being the least "instructional-like", interpreting the dimension with a bipolar label would be potentially misleading (Bohmann 2017: 326); hence only the positive end of the dimension will be labelled: 'Overt instructions and explanations'.

As in Chapter 6, for each dimension, linear mixed-effects models were computed to quantify the extent to which register, textbook proficiency level and the individual styles of textbook authors and publishers (as very approximately captured by the textbook series variable) contribute to each textbook texts' location on each principal component.[59] For the first dimension, the model featuring only Register as a fixed effect already explains 88% of the total variance in PC1 scores. Adding the nine textbook series as random, varying intercepts only very marginally increases the R^2 value to 90%, thus indicating that register is a remarkably strong predictor of PC1 scores, whereas the textbook series variable does not make a significant contribution to PC1 scores (see also plots of random effects in the Online Appendix 7.6). The anova-based comparison of the PC1 models showed that, whilst modelling Register*Level interactions provides a significantly better fit (as measured using AIC), only three interactions are significant at the level of $p < 0.01$. As shown in Table 66, these are: Instructional register with the textbook proficiency levels C, D and E.

Table 66: Summary of the model: lmer(PC1 ~ Register + Level + Level*Register + (1|Series))

Predictors	*Estimates*	*CI*	*p-value*
(Intercept) [Conversation, Level A]	-2.37	-2.59 – -2.15	**<0.001**
Register [Fiction]	1.61	1.36 – 1.87	**<0.001**
Register [Informative]	2.23	1.96 – 2.50	**<0.001**
Register [Instructional]	5.29	5.10 – 5.47	**<0.001**
Register [Personal]	0.48	0.08 – 0.88	**0.019**
Level [B]	-0.12	-0.30 – 0.05	0.167
Level [C]	0.12	-0.05 – 0.29	0.159
Level [D]	0.23	0.06 – 0.41	**0.01**
Level [E]	0.27	0.07 – 0.48	**0.01**
*Register [Fiction] * Level [B]*	0.18	-0.15 – 0.51	0.284
*Register [Informative] * Level [B]*	0.36	0.02 – 0.70	**0.038**
*Register [Instructional] * Level [B]*	-0.10	-0.35 – 0.15	0.434

[59] Detailed statistics and additional plots of all the models computed as part of this chapter can be found in the Online Appendix 7.6.

*Register [Personal] * Level [B]*	0.11	-0.39 – 0.61	0.671
*Register [Fiction] * Level [C]*	-0.25	-0.58 – 0.07	0.13
*Register [Informative] * Level [C]*	0.00	-0.32 – 0.31	0.993
*Register [Instructional] * Level [C]*	-0.39	-0.62 – -0.15	**0.001**
*Register [Personal] * Level [C]*	-0.22	-0.72 – 0.28	0.381
*Register [Fiction] * Level [D]*	-0.05	-0.38 – 0.27	0.739
*Register [Informative] * Level [D]*	-0.01	-0.33 – 0.31	0.946
*Register [Instructional] * Level [D]*	-0.47	-0.72 – -0.23	**<0.001**
*Register [Personal] * Level [D]*	-0.07	-0.60 – 0.46	0.8
*Register [Fiction] * Level [E]*	-0.24	-0.58 – 0.10	0.173
*Register [Informative] * Level [E]*	0.06	-0.29 – 0.40	0.747
*Register [Instructional] * Level [E]*	-0.50	-0.77 – -0.22	**<0.001**
*Register [Personal] * Level [E]*	-0.18	-0.74 – 0.38	0.527
	Random Effects		
σ^2	0.45		
$\tau_{00\ Series}$	0.07		
ICC	0.14		
N_{Series}	9		
Observations	1961		
Marginal R^2 / Conditional R^2	0.890 / 0.906		

Detailed inspection of the mean log *z*-scores of the linguistic features with high absolute loadings on PC1 (see Fig. 59 and Table 68) across the five textbook proficiency levels revealed that these significant Instructional*Level interactions are due to the number of imperative verbs (VIMP) featured in instructions and explanations progressively decreasing as textbook proficiency level increases, whilst the number of present tense verbs increases. Two reasons explain this. First, textbook instructions become more complex as textbook authors expect learners' proficiency in English to increase. This is illustrated in extracts (219) and (220), which stem from a level A and a level E textbook and which, on PC1, score 3.89 and 0.63 respectively. Second, secondary school beginner textbooks tend to include far fewer explanations in English, preferring to explain, e.g., grammatical concepts in the students' L1/school language. This means that the level A and B instructional texts of the TEC include fewer explanations than levels C, D and E textbooks. That said, whilst these three Instructional*Level interaction terms are significant and interpretable, the estimated differences in PC1 scores remain small and the marginal R^2 value of 89% and conditional R^2 of 90.6% of the model summarised in Table 66 make clear that, as compared to the model that only included Register as a fixed effect, the impact of textbook proficiency level on PC1 scores is only very marginal.

(219) **Identify** the people on the photograph. **Look** and **describe** what you can see. **Compare** the people. **Listen** and **describe** the characters' families. **Use** the genitive. <TEC: Piece of Cake 6e>

(220) Reactions
a) Describe what **happens** in the second half of the story (after line 43). How **do** the customers react? How **does** the narrator react?
b) **Do** you understand the way they **react**?
Short stories often **start** unusually ("medias in res" - right in the middle of the action) and **end** with a surprise. Look at "Deportation at breakfast" again: find these elements and say **why** they **are** important here. <TEC: Green Line 5>

The lack of overlap in the confidence intervals on Fig. 64 and the figures in Table 67 show that all of the register differences in estimated mean PC1 scores are significant, which confirms that PC1 distinguishes remarkably well between the different registers of the TEC data. Since PC1 accounts for 21.08% of the total variance in the TEC data, this confirms that much of the linguistic variation in Textbook English is register-driven.

Table 67: Estimated differences between mean PC1 scores for each TEC register pair (averaged across all textbook levels and series)

Contrast	*Estimate*	*SE*	*df*	*t-ratio*	*p-value*
Conversation - Fiction	-1.55	0.05	1962.95	-30.53	**<.001**
Conversation - Informative	-2.34	0.05	1960.92	-50.34	**<.001**
Conversation - Instructional	-4.99	0.04	1961.12	-125.14	**<.001**
Conversation – Personal Correspondence	-0.41	0.08	1958.34	-5.13	**<.001**
Fiction - Informative	-0.79	0.06	1962.12	-14.13	**<.001**
Fiction - Instructional	-3.44	0.05	1961.76	-69.17	**<.001**
Fiction – Personal Correspondence	1.15	0.08	1957.80	13.65	**<.001**
Informative - Instructional	-2.65	0.04	1956.92	-59.40	**<.001**
Informative – Personal Correspondence	1.93	0.08	1956.74	23.69	**<.001**
Instructional – Personal Correspondence	4.59	0.08	1956.63	58.82	**<.001**

Let us now turn to the linguistic features that load onto PC1 to find out which features contribute the largest register-based differences in the TEC. The most important ones are visualised in Fig. 59 and the full list of loadings is displayed in the first column of Table 68. As explained in 7.2.7, in the present framework and unlike in a classic MDA (Biber 1988: 93), every linguistic feature entered in the analysis loads onto each dimension, as opposed to only on the dimension to which they contribute most. As a result, in this model, all 61 linguistic features contribute, to a greater or lesser (sometimes extremely minimal!) extent, to PC1. Table 68 displays the feature loadings (eigenvalues) which shows the degree to which each feature correlates with each component. Positive values (in shades of yellow) contribute to high component scores, whilst negative ones (in shades of purple) contribute to low scores. These normalised weight values correspond to factor loadings in EFA: hence, to calculate a text's position along PC1, all the log *z*-scores of the 61 features of the text are multiplied by their corresponding PC1 loadings. Thus, if a text has a high average word length (AWL), its high (and positive) log *z*-score for

AWL will be multiplied by 0.22, which will contribute to placing this text high on the PC1 dimension. Should this text also feature no or very few verbal contractions verbs (CONT), its low (and negative) log *z*-score for contractions will be multiplied by -0.25, thus contributing to an even higher overall PC1 score. By contrast, a text consisting of mostly short words and featuring many contractions will likely score low on PC1. Very low absolute loadings are printed in light grey to indicate that these feature contributions most likely only represent noise and are therefore not considered in the interpretation of these dimensions of linguistic variation. However, they are not entirely removed from the table as a reminder that no threshold was applied in the calculation of the component scores.

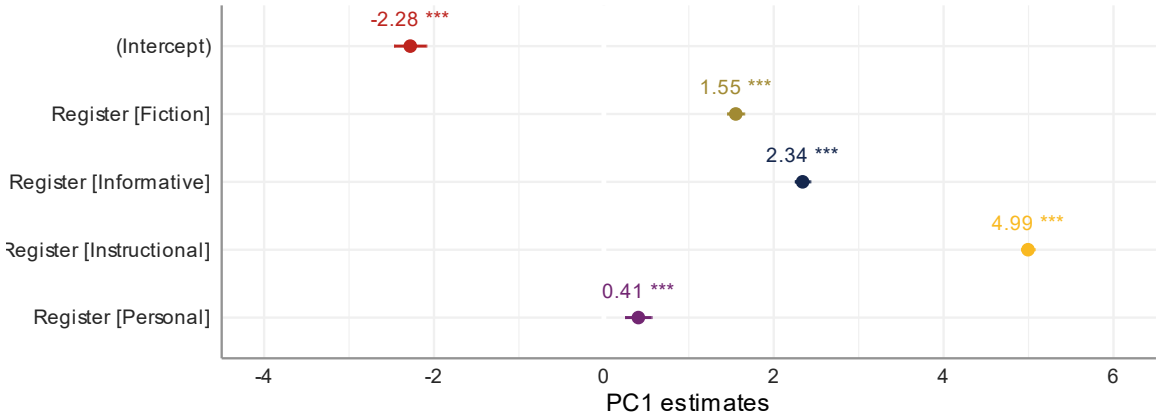


Fig. 64: Coefficient estimates and 95% confidence intervals of the fixed effects in the model: lmer(PC1 ~ Register + (1|Series)) (the intercept corresponds to the reference level: Register [Conversation])

Table 68: List of feature loadings (eigenvectors) on the four dimensions of the model of intra-textbook variation

	PC1	PC2	PC3	PC4
ACT	0.08	-0.11	0.04	-0.10
AMP	-0.12	-0.10	-0.11	0.01
ASPECT	0.10	-0.05	0.14	-0.01
AWL	0.22	-0.16	-0.12	-0.13
BEMA	-0.22	0.01	-0.22	0.02
CC	0.05	-0.21	-0.19	0.00
COMM	0.20	0.09	0.14	-0.04
COND	-0.01	-0.02	0.11	-0.25
CONT	-0.25	0.11	-0.03	-0.06
CUZ	-0.09	-0.13	-0.06	-0.02
DEMO	-0.12	0.08	0.03	-0.09
DMA	-0.20	0.14	-0.02	0.00
DOAUX	-0.01	0.20	0.06	-0.15
DT	0.12	0.00	0.31	-0.02
EMPH	-0.19	-0.02	0.06	-0.14
EX	-0.10	-0.05	-0.12	0.05

EXIST	-0.02	-0.15	-0.09	-0.09
FPP1P	-0.17	0.01	-0.07	0.00
FPP1S	-0.23	0.07	0.08	-0.01
FPUH	-0.16	0.15	-0.09	0.07
FREQ	-0.03	-0.05	0.01	-0.10
IN	0.17	-0.18	0.02	-0.08
JJAT	-0.06	-0.18	0.04	-0.21
JJPR	-0.17	-0.06	-0.12	-0.11
LD	0.16	-0.03	-0.26	-0.01
MDCA	-0.04	0.10	-0.18	-0.09
MDCO	-0.05	-0.10	0.22	0.01
MDWS	-0.07	-0.01	0.05	-0.17
MENTAL	0.14	0.13	0.12	-0.25
NCOMP	0.04	-0.05	-0.24	-0.15
NN	0.20	-0.09	-0.29	0.11
OCCUR	0.02	-0.18	0.03	0.02
PASS	-0.01	-0.22	-0.06	-0.06
PEAS	-0.06	-0.17	0.13	-0.13
PIT	-0.19	-0.04	-0.06	-0.06
PLACE	-0.16	-0.01	-0.07	0.09
POLITE	-0.14	0.13	-0.07	0.02
POS	-0.01	0.03	-0.04	0.16
PROG	-0.12	-0.02	0.11	0.00
QUAN	-0.15	-0.03	0.12	-0.19
QUPR	-0.10	-0.05	0.16	-0.11
RB	-0.19	-0.08	0.20	0.00
RP	0.00	-0.09	0.14	0.02
SPLIT	-0.11	-0.18	0.02	-0.16
SPP2	0.10	0.22	-0.01	-0.25
THATD	-0.05	0.04	0.16	-0.24
THRC	0.02	-0.11	-0.02	-0.18
THSC	-0.06	-0.17	0.07	-0.14
TIME	-0.13	-0.08	-0.01	0.06
TPP3P	-0.01	-0.16	-0.09	-0.02
TPP3S	-0.06	-0.11	0.13	0.30
TTR	-0.04	-0.26	-0.05	-0.01
VBD	-0.08	-0.20	0.23	0.30
VBG	0.04	-0.18	0.00	-0.22
VBN	0.03	-0.18	-0.07	-0.04
VIMP	0.25	0.15	0.04	-0.08
VPRT	-0.16	0.05	-0.32	-0.22
WHQU	0.11	0.23	0.00	-0.09
WHSC	0.11	-0.11	0.03	-0.15
XX0	-0.22	0.03	0.06	-0.06
YNQU	-0.03	0.23	0.00	-0.08

As shown in Fig. 56 and Fig. 58, the upper end of the PC1 scale, roughly between 2 and 4, is entirely reserved for instructional texts. We can therefore expect the linguistic features that load positively on PC1 to be highly frequent in textbook instructions and explanations. Indeed, these include imperative verbs (VIMP), the semantic categories of communication and mental verbs (COMM and MENTAL) and WH-questions (WHQU), e.g., (215). Other features contributing to high PC1 scores are more akin to the negative, 'informational' end of Biber's (1988) Dimension 1. These include longer words (AWL), nouns (NN), prepositions (IN), and a high ratio of content to function words (LDE), all of which are associated with impersonal and informational writing.

By contrast, the linguistic features with the most negative loadings on PC1 are associated with spontaneous, interactional production: e.g., contractions (CONT), first person singular and *it* pronouns (FFP1S and PIT), negation (XX0), discourse markers (DMA), emphatics (EMPH), fillers and interjections (FPUH), and demonstrative pronouns (DEMO), e.g. (221). These features very much echo the upper, interactional, end of Biber's (1988) Dimension 1 (see Table 40).

(221) **Hi**, Amy.
Hi, you two.
Hello. **What's so** funny? Nothing - honestly. **Well**, what were you **talking** about? You**'ve** got big wide grins on your faces!
Oh, **this** and **that**. You **know**, **just chatting**. **We** were **talking** about thriller films. **We're thinking** of watching one. **Want** to join **us**?
Yeah, count **me** in. Sure. **I** have**n't** seen a good film for far too long. **Got** anything in mind?
Well, **there** was one film **we** were **thinking** about... But **I've** seen **it** - and **anyway**, **it'd** be far too scary for you two!
Do you **want** to bet? **There's** never been a horror film that **I** did**n't** watch all the way through.
Take **it** easy, Nick - **I** think she**'s pulling** your leg!
Oh. Right. Sorry! <TEC: English in Mind 4>

7.3.1.2 Variation on Dimension 2: 'Involved vs. Informational Production'

The second dimension of variation in the TEC data (PC2) accounts for 14.16% of the total variance. As has been argued in 7.3.1, it shows a high degree of overlap with the linguistic features that contribute to both ends of Biber's Dimension 1 and will therefore also be labelled: 'Involved vs. Informational Production'.

A first mixed-effects model with random intercepts for each textbook series and only Register as a fixed effect explained 56% of the variance in scores on this second dimension, whereas the full model adding Register*Level interactions explained 72%. This indicates that linguistic variation along this second dimension is driven by both register and the proficiency level of the textbooks. As with the first dimension, a

comparison of the models showed that textbook series has no significant effect on the position of textbook texts along this dimension. As a result, only the estimated coefficients of the fixed effects of the full PC2 model are visualised in Fig. 65. The colours correspond to the textbook registers as already introduced in the previous plots of this chapter. The intercept represents Textbook Conversation Level A texts, and the coefficients are interpreted just like in the model summary tables (e.g., Table 66).

The register cline on this second dimension bears strong similarities to Biber's (1988) Dimension 1, 'Involved vs. Informational Production' (see Fig. 29), which, in various forms, has emerged as the strongest dimension of linguistic variation in many MDAs, across a wide range of domains and languages (Biber 2014). As shown on Fig. 56 and Fig. 58, Textbook Conversation texts are mostly clustered at the upper end of the cline. In addition, Fig. 65 indicates a clear proficiency-level effect – with more advanced textbook conversations scoring lower on average. The lower end of the register cline is dominated by Textbook Informative texts which all have negative PC2 scores. Personal correspondence and fiction are, once again, situated in the middle. The overlapping confidence intervals on Fig. 65 show that the two textbook registers cannot, on average, be reliably differentiated on this second dimension of linguistic variation.

As we would expect given that linguistic variation is multi-dimensional, we find in Fig. 59 and in the table of feature loadings (Table 68) that some of the linguistic features with positive loadings on PC1 also contribute to positive values on PC2: these include imperatives, mental and communication verbs – which are known to be particularly strongly associated with instructional texts, e.g., (215). In addition, we find WH- and *yes/no*-questions (WHQU and YNQU) and second-person references (SPP2), which are also found in many textbook instructions and task descriptions. Consequently, many instructional texts score relatively high on PC2, e.g., (222). In contrast to the first dimension, on this second dimension these features are also associated with linguistic features typical of spontaneous conversation and interactional language, in particular with fillers and interjections (FPUH), discourse (DMA) and politeness markers (POLITE), the modal *can* (MDCA) and singular first-person references (FPP1S). As a result, the highest PC2 scores are achieved by textbook dialogues such as (217), which feature rapid question and answer exchanges. Some of the highest PC2 scores are found in textbook dialogues that model classroom interactions. These incorporate spoken instructional language which sometimes includes some of the instructions that accompany the textbooks' tasks and exercises verbatim, e.g., compare extracts (222) and (223).

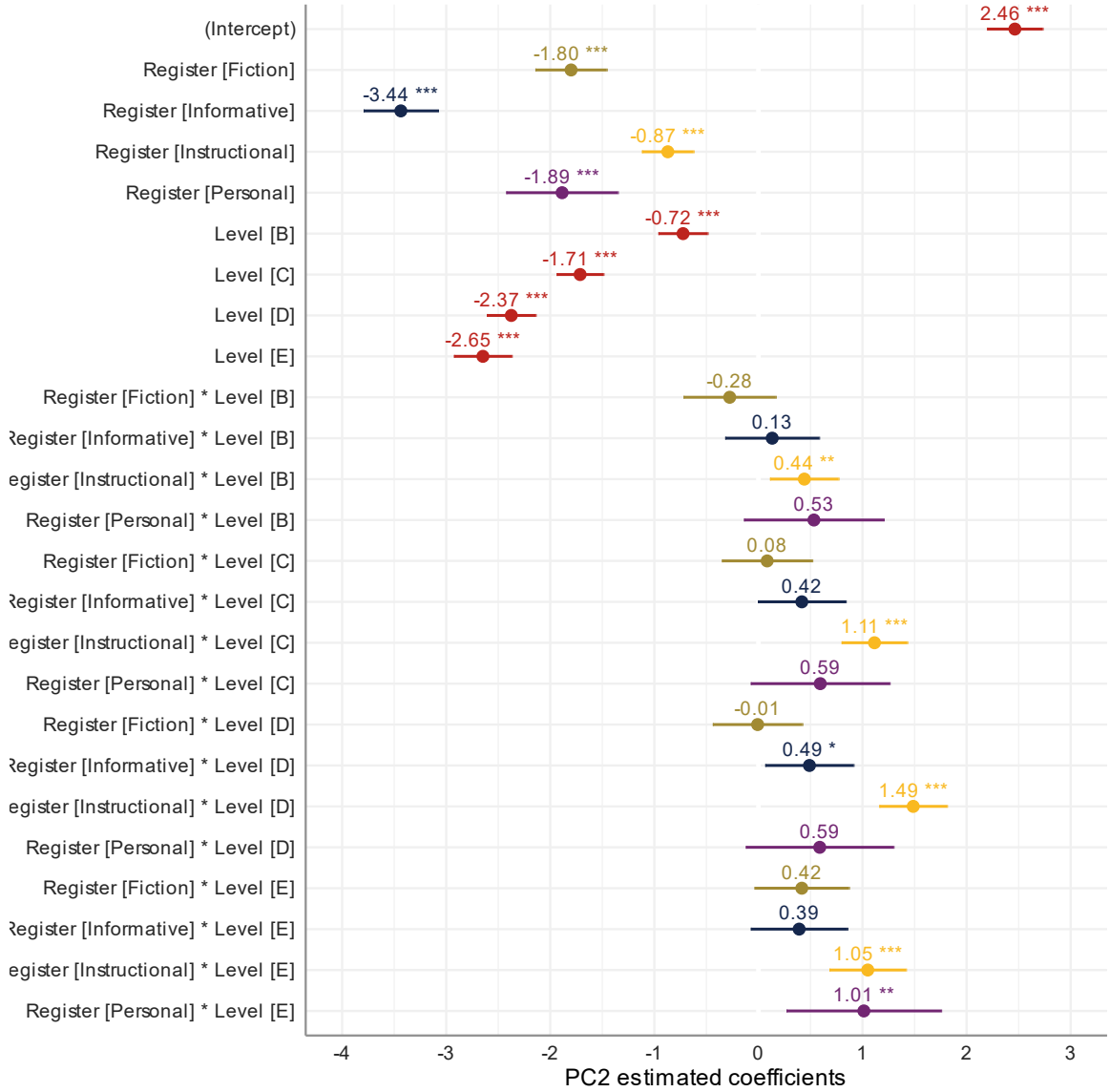


Fig. 65: Coefficient estimates and 95% confidence intervals of the fixed effects in the model: lmer(PC2 ~ Register + Level + Register*Level + (1|Series)). The reference levels are Register [Conversation] and Level [A].

(222) Nadia is going on holiday A. **Read** her email to her friend. **Who** is going with her on holiday**? What does** she promise to do after her holiday**?**
(T) **Imagine you** are going on holiday.
B. **Write** an email to an English-speaking friend and **tell** him/her what **you**'re going to do. **Use** the information and Nadia's email to help **you**.
<TEC: English in Mind 2>

(223) Good morning, everyone! **Right**, **sit** down! Paul, **can you** show us where Scotland is?
OK, Miss.
Excellent. **Please take your** workbooks out. Now, **please open your** workbooks at page 12. Yes?
Sorry, Miss, my workbook is at home.
Can you work with Paul, **please**?
OK. <TEC: Access G 1>

The feature with the strongest negative contribution to this second dimension is type/token ratio (TTR) – a feature which, in Biber's (1988) model, is also strongly associated with the 'informational' end of the 'Involved vs. Informational Production' dimension. In addition, the lower end of the dimension is characterised by further lexico-grammatical features typical of structurally more complex, meticulously drafted written production such as passives (PASS), coordinating conjunctions (CC), non-finite verb forms ending in *-ing* and *-ed* (VBG and VBN), split auxiliaries (SPLIT) and *that*-subordinate clauses (THSC), e.g., (224). Like the informational end of Biber's (1988) Dimension 1, prepositions (IN), longer words (AWL), and attributive adjectives (JJAT) are also associated with the negative end of PC2 (see Table 68).

(224) Although **books** are still popular **with** teenagers, most **of** them spend more of their **leisure time staring at** their **phone than reading** a paperback. And the more versatile **phones** become, the more **reasons young people** have **for looking at** them. **In response to** this **trend**, some **smart**, **young authors** have changed the **way** they write. **Instead of publishing** a **whole book at** once, they produce very **short chapters**, which they send once a **week to** their **readers by text message**. Some even claim **that** this **style of writing** represents a **new literary genre**: the '**cell phone novel**'. <TEC: Solutions intermediate>

Fig. 65 shows a clear pattern of decreasing PC2 scores as the proficiency level of textbooks increases. A cursory look at the linguistic features with negative loadings on PC2 (see Table 68) suffices to understand why: the majority of these features are not introduced until the second or third year of EFL instruction. The extent to which these features are intrinsically linked to specific registers determines how large the shift to the negative end of the PC2 scale is, as learners are progressively introduced to these features in the more advanced textbooks.

Thus, we find that the median PC2 score for beginner textbook fiction texts is 1.09 (MAD = 0.59) but, as soon as the past tense (VBD) is introduced in level B textbooks, the PC2 scores for fiction texts drop to a median of -0.06 (MAD = 0.74) and then further to -0.82 (MAD = 0.76) in level C fiction. The other features that make significant contributions to this negative shift include higher type/token ratio and lexical density (LD), longer words (AWL), and higher normalised frequencies of the perfect aspect (PEAS), passives (PASS), *could* as a modal (MDCO), occurrence and existential verbs (OCCUR and EXIST), prepositions (IN), attributive adjectives (JJAT), and *that*-subordinate clauses (THSC), e.g., (225). Fig. 66, however, suggests that, on average, the linguistic features that contribute to PC2 are used to a similar extent in level D and E fiction texts. Indeed, apart from type/token ratio, which can be expected to continue to grow as learners become more proficient in English, all the aforementioned features that contribute to the negative end of PC2, e.g., perfect aspect, passives, *could*, etc., can be expected to have been taught by the fourth year of secondary school EFL instruction.

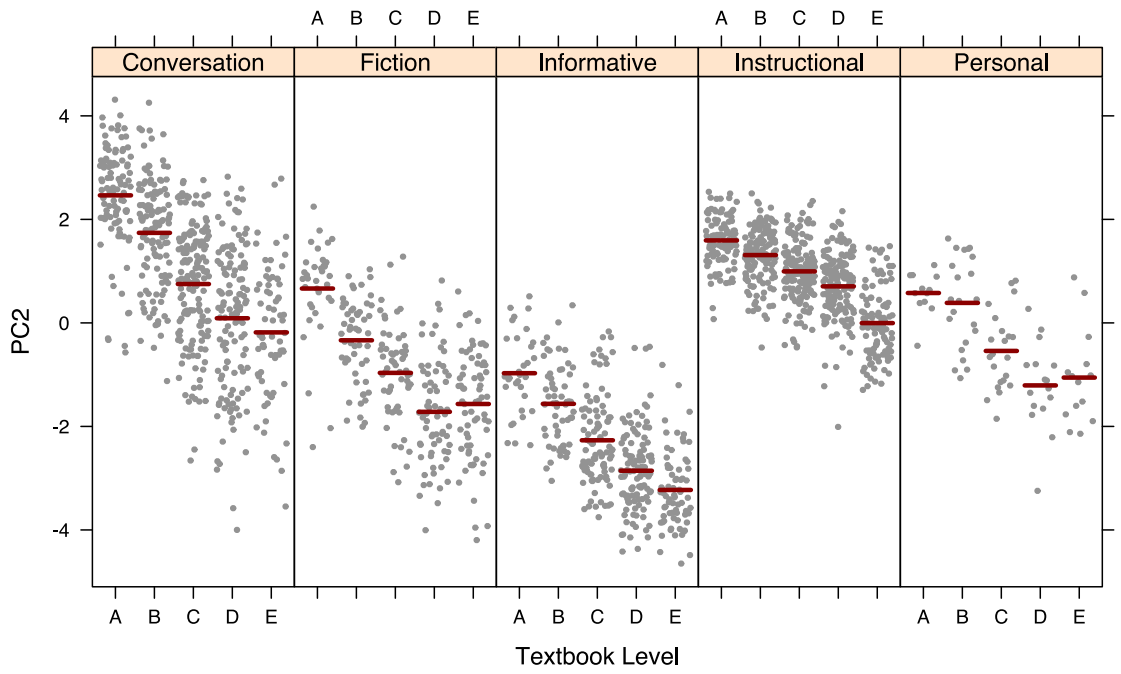


Fig. 66: Estimated PC2 scores across each register and the five textbook proficiency levels

(225) "If I just **had** that knife **in** my hand, I **could** ..." Suddenly a smile **lit** up his face as he **thought of** a plan. He **took** hold **of** his fishing line. The pole **was** still **in** the canoe so he **could** pull the boat **towards** him, get the knife, cut the roots and get free. Carefully, he **began to** pull **on** the line and **felt** the canoe start **to** move. Then, just as he **started to** so hope **that** his plan might be working, he **heard** a **quiet** splash as his fishing pole **fell** out **of** the canoe **into** the water. The canoe **stopped** moving. […] **While** he **had watched** the alligator, his canoe **had drifted** up **behind** him and **hit** him **in** the back. <TEC: Access G 4>

For the informative and conversation texts, the shift towards the negative end of the PC2 scale from level A to E is more or less linear (see Fig. 66). This pattern is to be expected for informative texts: as for the narrative texts, textbook authors are necessarily restricted in their choice of grammatical features given that learners have only been introduced to a limited number of grammatical features and basic vocabulary. Many of the negative loading features on PC2 are typical of informational writing so that it comes as no surprise that, on average, the informative texts of the more advanced textbooks score lowest on this dimension, e.g., (226).

(226) A kiss is ambiguous **at** the best **of** times, **signifying** anything **from** friendliness **to** desire, deference **to** insult. Kissing - **on** the lips, originally - **was**, **in fact**, a common form **of** social greeting **in** Britain **from** Roman times **at least until** the 1700s, **when** the potential **for** misinterpretation **led to** its disappearance. Abroad, of course, they'**ve never really abandoned** the gesture, **although** the rules **governing** its use **are sometimes exceedingly complicated**. **In** France, **for example**, anything **between** one **and** four kisses can be acceptable **depending on** who you are,

who you're kissing, how well the two **of** you know each other **and** exactly where you both happen **to** be **in** France. There are so many variables **that** even French people **within** the same region confess **to being** confused. <TEC: New Bridges 2[de]>

By contrast, the strong interactions between the conversation register and the proficiency levels of the textbooks constitute a more puzzling finding because few of the features with negative loadings on PC2 can be said to be typical of real-life conversation; yet, on average, advanced textbooks nevertheless feature considerably more of these in their representations of spoken language than beginner textbooks (see Fig. 66). The median PC2 value of level A textbook conversations is 2.60 (MAD = 0.76), with many beginner conversations scoring considerably higher, e.g., (217) and (223) with PC2 scores of 4.08 and 3.53 respectively. However, median PC2 scores progressively drop as textbooks become more advanced, reaching -0.16 (MAD = 1.33) for level E textbook conversations. This echoes the results of the additive MDA based on Biber's (1988) Dimension 1 (see 6.3.1.1 and 6.3.2.1). Advanced textbook dialogues with low PC2 scores tend to feature much higher type/token ratios, longer words, more passives, past tense and perfect verbs, split auxiliaries, coordinating conjunctions, nouns, prepositions, and subordinate clauses, e.g., (227)–(228).

(227) We're here **at** the **BBC Radio**'s annual **Teen Awards at Wembley**. **As** I'm sure many **of** our **listeners** know, the **prizes are awarded to** the **year**'s best **vloggers**, **sport and music stars and to teenage heroes who have inspired** everyone! Best **of** all, they **have been voted for by Britain**'s **teenagers**! So let's find out **what** the **fans** here **thought of** the **show**. OK, what **did** you think **was** the best **moment of** the **afternoon**?
Well, **for** me, it has to be **when Jack G got** his **award for** standing up **to** bullying. If I'**d been** him, I wouldn't **have had** the **courage to** start a **campaign against** the **bullies in** my **school**, so I really admire him **for** doing that. <TEC: Solutions intermediate plus>

(228) The **state** is **also** very active **in** limiting **air pollution**. **But** what **about traffic in** places like L.A.?
Yes, the **traffic in Los Angeles** is a huge problem **which has existed since** the **arrival of** the **automobile in** the late 1800s. **Until then people were transported in L.A. by** streetcar. **Once** the **automobile arrived**, **people came** to love the **freedom** it **offered**. It **allowed** them **to** move far **from** the **center of L.A. and still** be able **to** reach **downtown** - something the **streetcars** couldn't offer. <TEC: Green Line New 5>

The pattern observed in the instructional register on Fig. 66 is in line with that observed in the previous section (7.3.1.1). Linguistically, it is driven by more complex sentence structures in the explanations of more advanced textbooks which are characterised by higher frequencies of subordinate clauses (THSC and WHSC) and coordinating conjunctions (CC) – three features which contribute to lower PC2 scores. This leads to the modest, but nevertheless significant, interaction effects between the instructional register and textbook proficiency level reported in the mixed-effects

model of PC2 scores (see Fig. 65, Fig. 66 and, for the full results of the model, the Online Appendix 7.6).

7.3.1.3 Variation on Dimension 3: 'Present/factual vs. Past/speculative'

The third dimension that emerges from this intra-textbook PCA accounts for just 5.31% of the variance. As already identified in 7.3.1, this dimension of variation appears to involve both register-based and proficiency level-based linguistic variation (see Fig. 60 and Fig. 62). This visual observation was confirmed in the mixed effects models computed to model PC2 scores: a comparison of models for this dimension shows that register alone explains 27% of the variance in PC3 scores; the proficiency level variable alone accounts for 12%, but by modelling the interactions between the two variables 44% of the variance in PC3 can be explained. The fixed effects coefficient estimates of the full model that includes these interactions are displayed in Fig. 67.

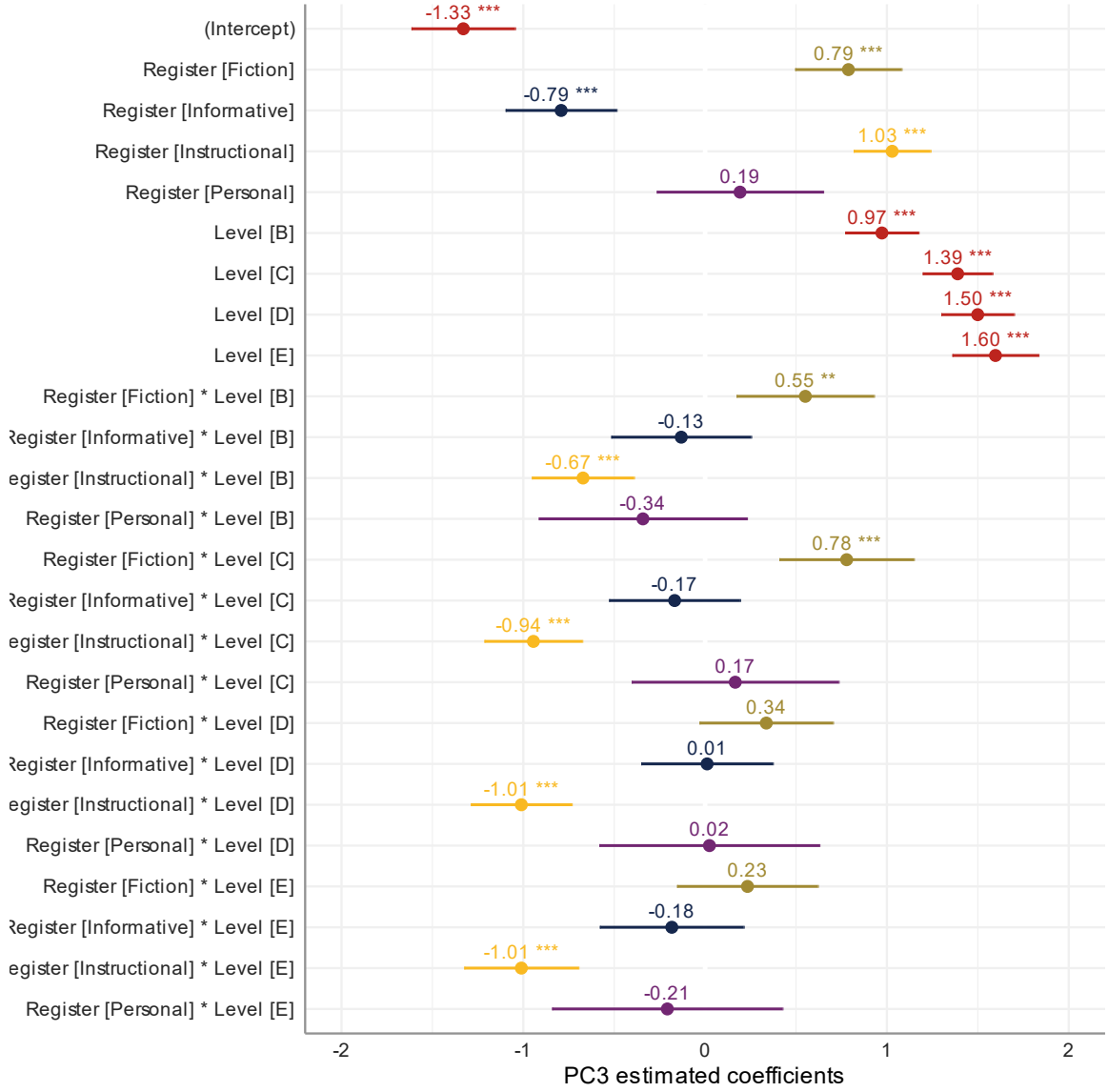


Fig. 67: Coefficient estimates and 95% confidence intervals of the fixed effects in the model: lmer(PC3 ~ Register + Level + Register*Level + (1|Series). The intercept corresponds to the reference levels Register [Conversation] and Level [A].

The effects of the interactions are also illustrated in Fig. 68. Across all five registers, the largest jump in PC3 scores is observed between Level A and Level B textbook texts. It is largely driven by the near absence of past tense verbs in Level A textbooks contrasted by the highly frequent use of the past in Levels B and C textbooks as learners are taught to comprehend and produce this grammatical feature (see 6.3.1.2 for a similar observation on Biber's Dimension 2 including examples). Other strongly loading features on PC3 whose frequencies are also strongly mediated by the proficiency level of the textbooks include the perfect aspect (PEAS), the use of the modal *could* (MDCO), and conditional subordination (COND) – all three of which make positive contributions to PC3 scores.

Interestingly, aside from the present tense variable (VPRT), the frequencies of all the other strong contributors to the negative end of this dimension are not correlated with textbook proficiency and hence do not contribute to the effects visualised in Fig. 68. Instead, this cluster of negative-loading features – which includes the total frequency of nouns (NN), lexical density (LD) and noun-noun compounds (NCOMP) – describes the nominal style of the negative end of this dimension. This is a characteristic of Textbook English that is stable across all proficiency levels. The multitude of effects at play in this dimension make it difficult to interpret but, for now, this dimension is tentatively labelled 'Present/factual vs. Past/speculative'.

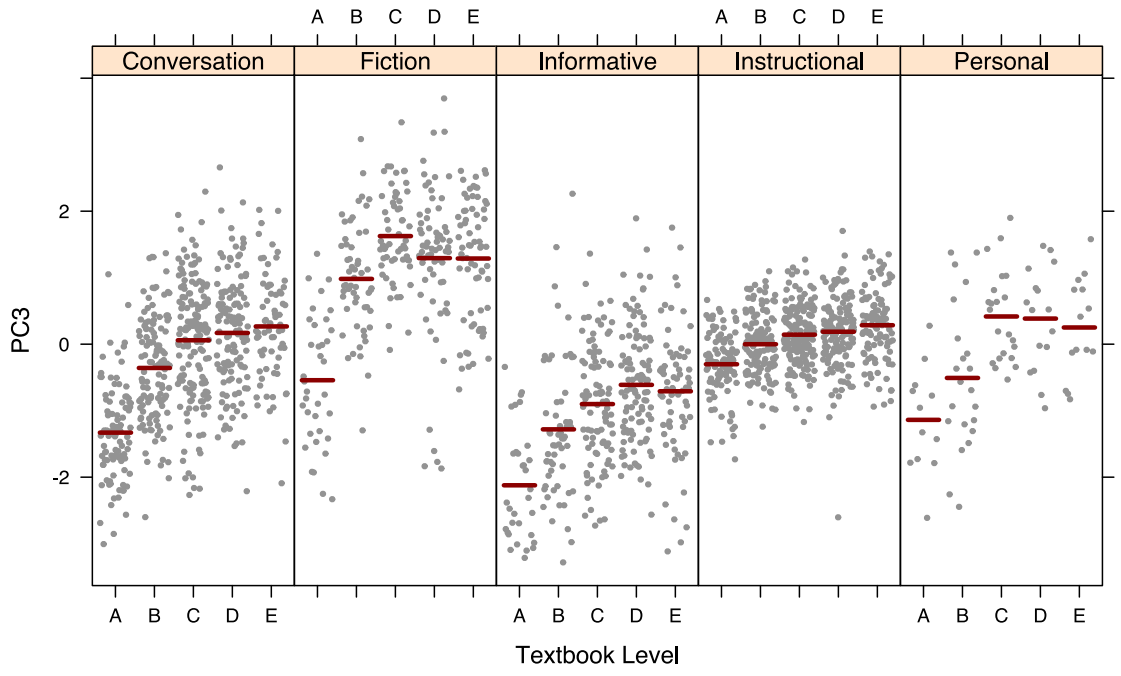


Fig. 68: Estimated PC3 scores across each register and the five textbook proficiency levels

7.3.1.4 Variation on Dimension 4: 'Clausal complexity'

Similarly to PC3, the fourth dimension that emerged from this intra-textbook PCA only accounts for 4.50% of the variance in the texts of the TEC. It, too, reflects both

register- and proficiency-level-based variation – this time to almost equal degrees: register alone explains 20% of the variance in PC3 scores, whilst the five levels of the proficiency level variable account for 19%. The full mixed effects model involving the interactions between the register and proficiency levels, however, explains a total of 43% of the variance in PC4 scores. The estimated coefficients of the fixed effects of this model are plotted in Fig. 69.

At first glance, the highest loading features on this fourth dimension might suggest a second 'narrative' dimension since some of these features (e.g., past tense verbs and third-person reference) are shared with the positive-loading features of the third dimension (see Fig. 61). However, this dimension appears to be much more driven by the interaction effects of proficiency level and register. These effects, predicted by the model summarised in Fig. 69, are illustrated in Fig. 70. Their direction is somewhat surprising given that the second highest loading feature on PC4 is past tense verbs (VBD, see Table 68) and, whilst we have seen in Fig. 47 in 6.3.2.2 that the occurrence of past tense verbs drastically increases from level B textbook onwards, Fig. 70 show that PC4 scores actually decrease as the proficiency level of the textbook increases.

Thus, this proficiency-level-based effect must be driven by other lexico-grammatical phenomena that contribute to negative scores on this dimension and which are gradually introduced in the later years of secondary school EFL teaching. These include *if*-conditionals that lead to an increase in conditional subordinators (COND) and the modals *will* and *should* (MDWS), e.g., (230). Furthermore, this dimension points to linguistic variation that, at the upper end, is characterised by narrative discourse that relies foremost on individual words for brief descriptions (e.g., adverbs of time and place and use of the *there is/are* construction, see (229)), whilst, as attested by the negative loadings of features such as THATD, VBG, THRC, WHSC and THSC (see Table 68), the lower end of this fourth dimension is characterised by more complex sentence structures that, in fiction, allow for more detailed descriptive passages and in conversation, personal correspondence, instructional and informative texts, convey more sophisticated explanations, opinions or arguments, e.g., (230). Hence, this dimension can be summarised as representing a complexity cline with the simplest constructions clustering at the upper end of the dimension and the most complex at the bottom. The label 'Clausal complexity' was therefore chosen to describe this fourth dimension.

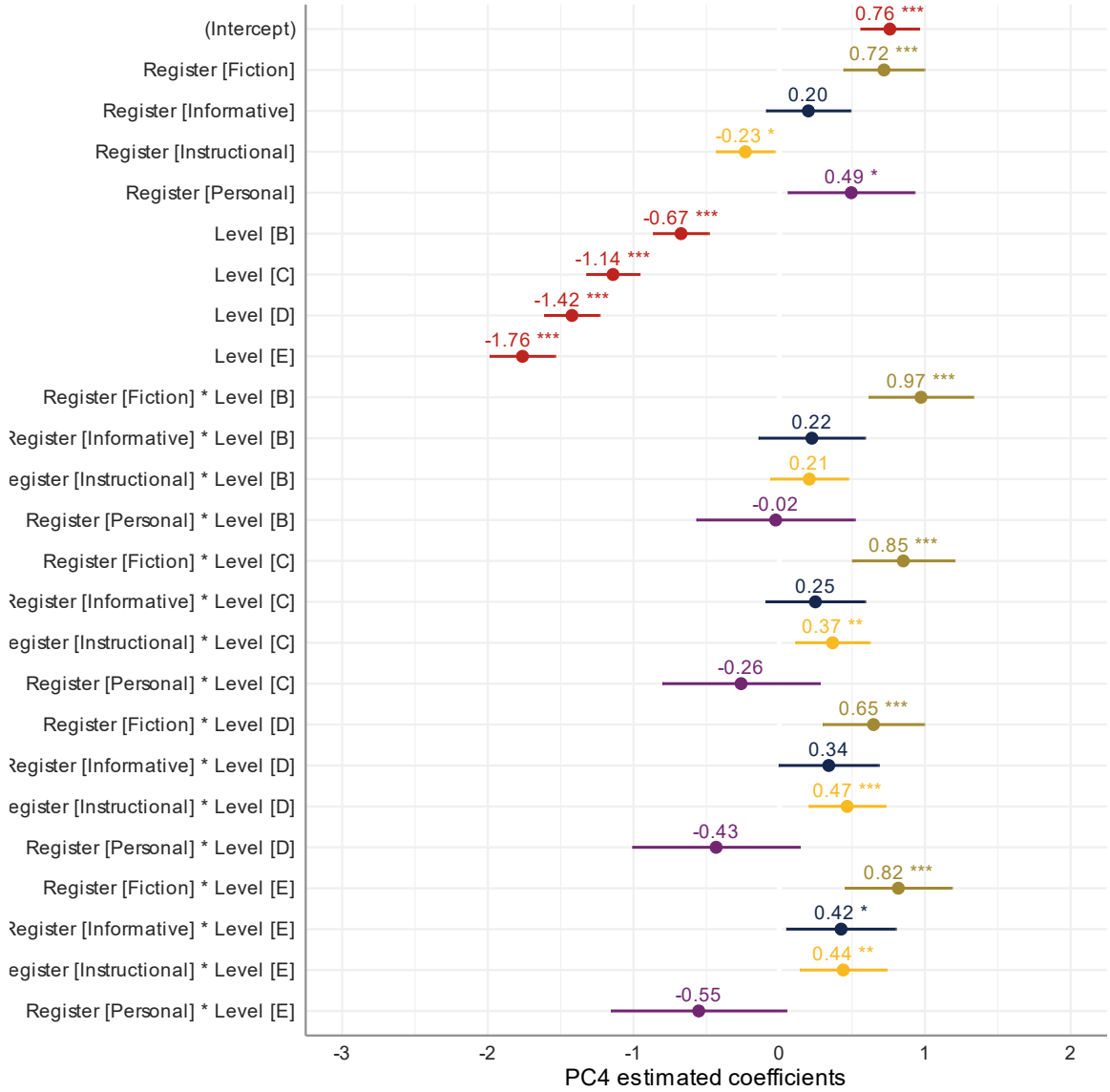


Fig. 69: Coefficient estimates and 95% confidence intervals of the fixed effects in the model: lmer(PC4 ~ Register + Level + Register*Level + (1|Series)). The reference levels are Register [Conversation] and Level [A].

(229) In December 1980, ex-Beatle John Lennon and his wife Yoko Ono **were** in New York. In the afternoon, they **were** on their way to a recording studio to work on a new song. **There was** an American called Mark Chapman in the street. In his hand, **there was** a piece of paper and a pen. 'Mr Lennon,' he **said**. 'Can I have your autograph?' John Lennon **signed** his name and Chapman **went** away. In the evening, John and Yoko **were** in front of their apartment building. **There was** a man at the door. It **was** Mark Chapman. This time, there wasn't a pen in his hand, but a gun. 'Mr Lennon!' he said. Suddenly, **there were** five shots and John Lennon **was** dead. <TEC: English in Mind Starter>

(230) On the other hand, opponents of nuclear weapons argue **that** it **will** not be long before some countries develop a defence system **that** makes them immune to nuclear attack, while still **being** able to launch a nuclear offensive of their own. When this happens, the world **will** suddenly become an extremely dangerous place. Furthermore, they point to the huge

cost of the weapons and say **that** the world would be a better and safer place **if** the money were spent on health and education. <TEC: Solutions Intermediate plus>

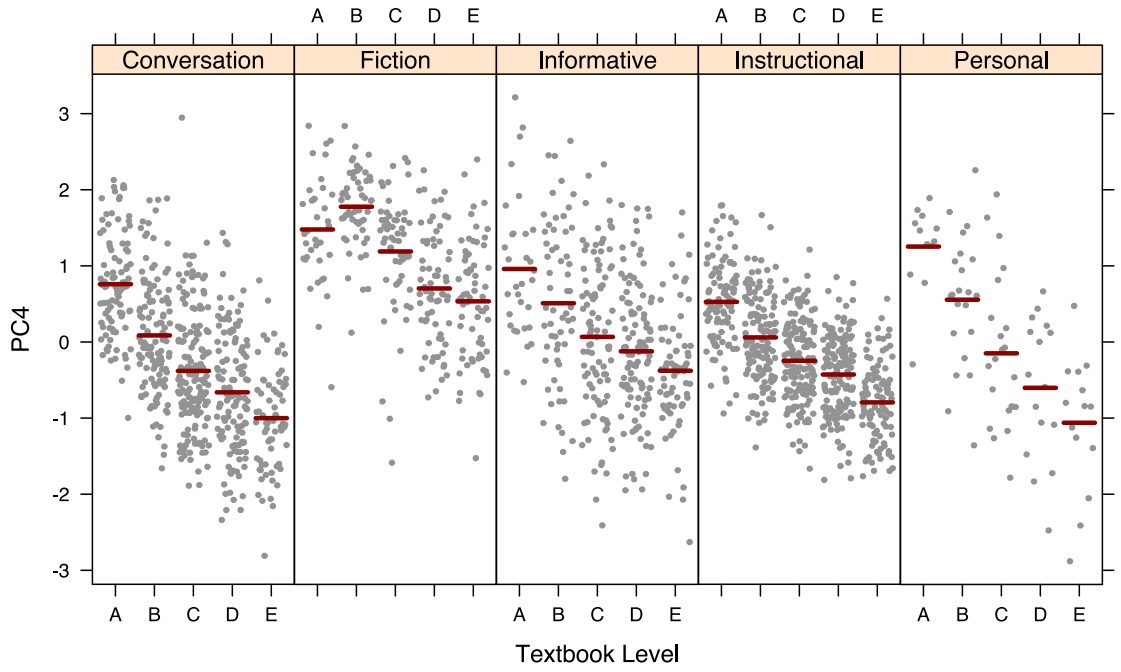


Fig. 70: Estimated PC4 scores across each register and the five textbook proficiency levels

7.3.2 PCA of Textbook English vs. 'real-life' English

This section presents the results of a PCA-based MDA that attempts to capture the similarities and differences between the texts of the Conversation, Fiction and Informative subcorpora of the TEC (see 3.3.1.4) and the three corresponding reference corpora outlined in 3.3.2: the Spoken BNC2014, the Youth Fiction corpus and the Informative Texts for Teenagers Corpus (Teens Info). Methodologically, it follows the same procedure as for the intra-textbook variation PCA reported on in 7.3.1.

First, the exclusion of linguistic features that are absent from more than two-thirds of the texts of the TEC and reference corpora together led to the merging of the ABLE and JJPR categories and the PGET and PASS (as in 7.3.1), as well as the removal of the EMO, HST, PRP and URL features. Due to otherwise very low final communalities, the singular third-person (TPP3S) and the plural third-person referent categories (TPP3P) were merged into a single TPP3 category. For the same reason, the adverbs of frequency (FREQ) and time (TIME) were also combined into a more general FQTI category. The normalised counts of all the features were then standardised before plotting their distributions (see Online Appendix 7.5). Next, 115 outlier texts (= 2.28% of the texts) were identified on the basis of excessively high z-scores. The majority of these outlier texts ($n = 74$) stem from the Info Teens corpus. Many of these outliers are articles composed of bullet point lists featuring very few

finite verbs, leading to improbably high counts of some of the linguistic features normalised per finite verb phrase (FVP) (see 7.2.4). These texts were therefore removed from the dataset to be entered in the following PCAs to avoid them exerting undue influence on the model and inflating or distorting differences between texts (see, e.g., Le Foll 2021a: 110–113 on the consequences on leaving such texts in such analyses). Here, too, signed log transformation was applied to tame some of the highly skewed feature distributions (see Online Appendix 7.5 for details of the procedure and plots of the distributions before and after transformation). The dataset was checked for collinearities and excessive correlations with the help of a correlation matrix display (see Fig. 53 and Online Appendix 7.5). This led to the exclusion of the present tense feature (VPRT) because, with a correlation of -0.97, its normalised counts per FVP are almost the perfect mirror image of past tense counts per FVP (which, given that finite verbs can either be tagged as past tense, present tense, or modal does not come as a surprise). As a result of the above steps, none of the remaining features returned individual MSA values < 0.5 or final communalities of < 0.2 so that the final data matrix consisted of the signed log transformed standardised normalised counts of 72 features in 4,980 texts. Its overall KMO value of 0.95 suggests that it is "marvellously" suitable for factor analysis (Kaiser & Rice 1974: 112). The scree plot of eigenvalues suggested a four-component solution (see Fig. 71). The cumulative proportion of variance explained by these four principal components is 47.15%.

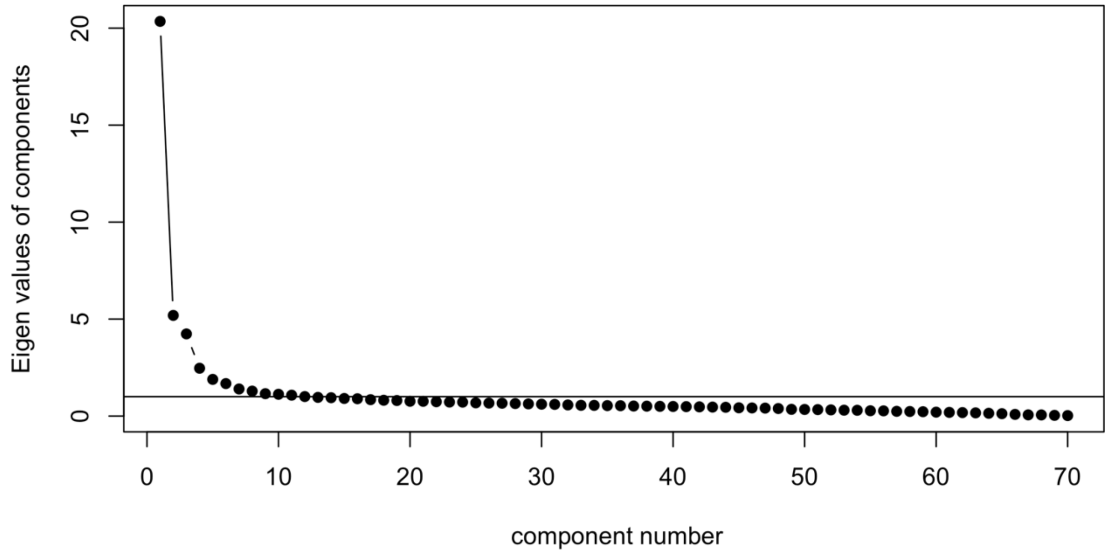


Fig. 71: Scree plot of the eigenvalues of the PCs for the Textbook English vs. 'real-life' English PCA

As with the exploration of intra-textbook linguistic variation in 7.3.1, the retained components were first explored using 3-D projections of the position of the texts along the four dimensions. The projection of the first three principal components (see Fig. 72) reveals three very well-defined clusters for the three reference corpora (in red, bright green and dark blue) – with hardly any overlap. By contrast, the three TEC

subcorpora cluster across much larger areas of the 3-D plot, among them the intercept of the plot, which is an area in which hardly any of the reference texts are found.

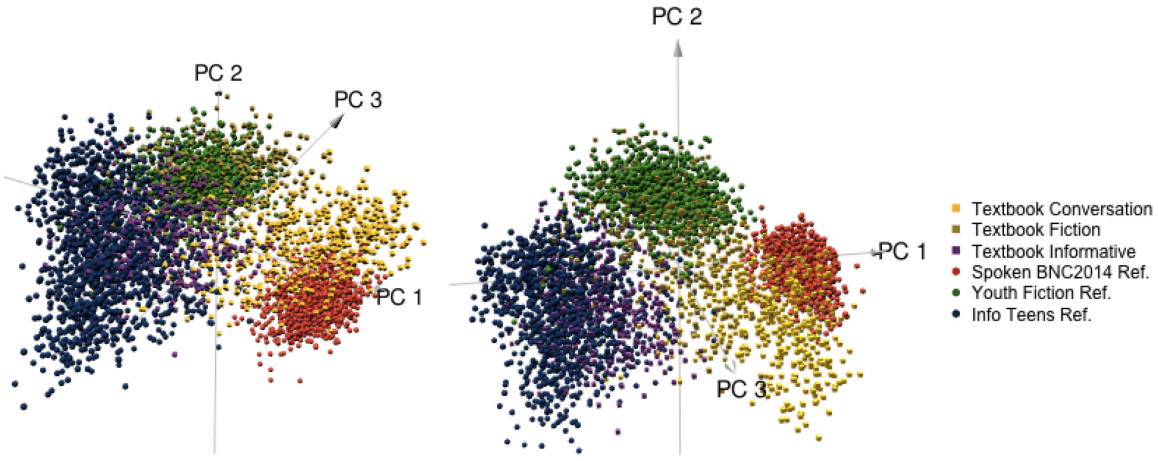


Fig. 72: Snapshots from the 3-D representation of texts along PC1-PC3

This pattern can also be observed in the scatterplot matrix of all combinations of the four dimensions in Fig. 73. At first sight, it would appear that Textbook Fiction texts and those of the Youth Fiction corpus share many similarities on all dimensions. This is in stark contrast to the Textbook Conversation and Spoken BNC2014 clusters, which overlap very little on all four dimensions. In addition, Fig. 73 indicates that there is a lot of internal variation within the Info Teens corpus.

Together, the first two principal components explain 37.85% of the total variance. As shown in more detail in Fig. 74, both these dimensions of linguistic variation appear to capture differences between the three textbook registers and their corresponding reference corpora. The corresponding graph of features (Fig. 75) shows that these first two dimensions share many similarities with Biber's (1988) first two dimensions. The first dimension (PC1) places highly interactional, spontaneous speech at the upper end, whilst informationally dense, edited texts score lowest. It will therefore be labelled 'Spontaneous interactional vs. Edited informational'. On both ends of the dimension, many of the strongest contributing features overlap with those that also have the highest absolute loadings on Biber's (1988) Dimension 1, e.g., verbal contractions (CONT) and discourse markers (DMA) at the positive pole and nouns (NN), longer average word length (AWL), prepositions (IN) and higher lexical diversity (TTR) at the negative pole.

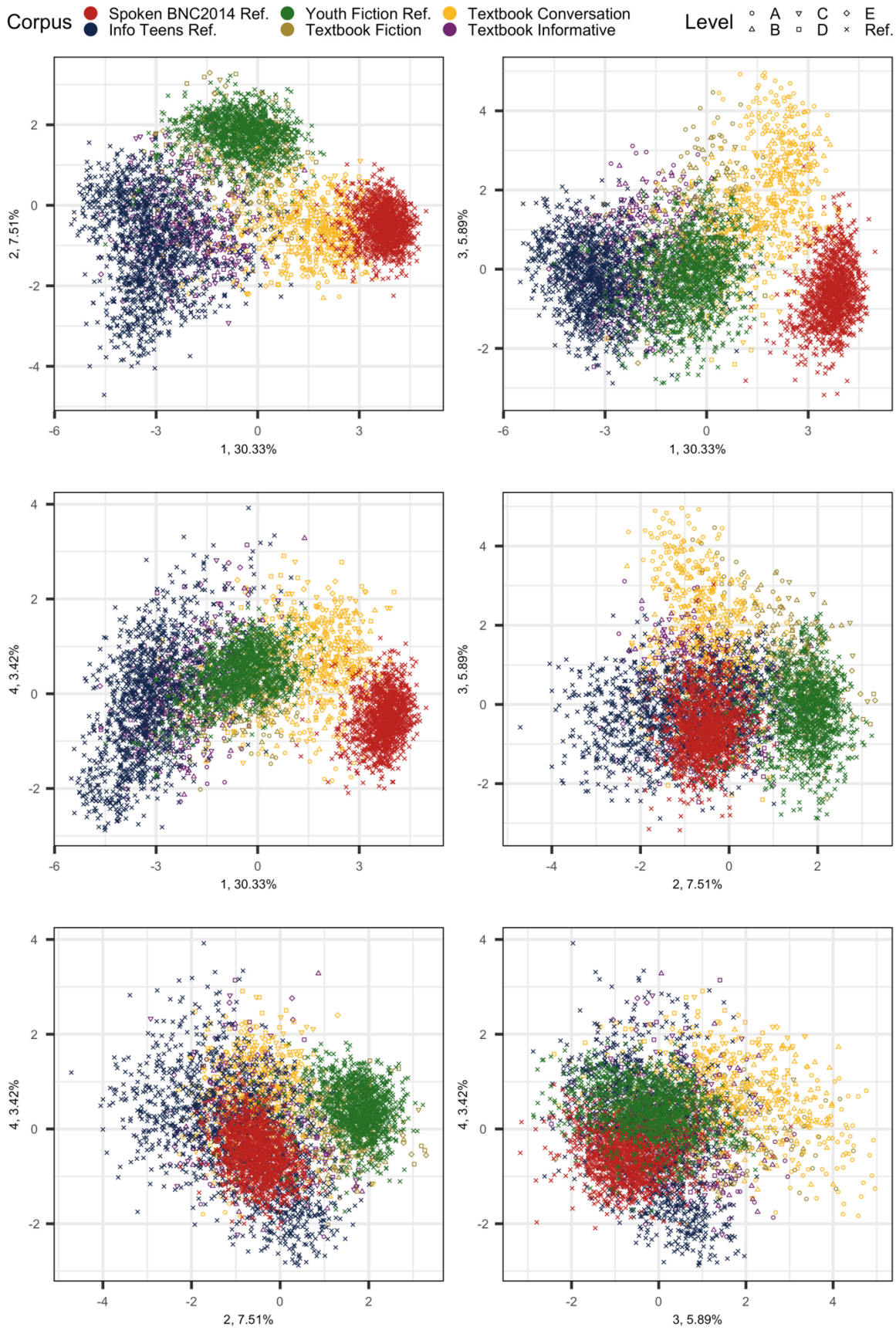


Fig. 73: Scatterplot matrix of combinations of the four dimensions of the model of Textbook English vs. 'real-life' English (the number before the comma on each axis label shows which principal component is plotted on that axis; this is followed by the percentage of the total variance explained by that particular component)

The second dimension (illustrated on the *y*-axes of Fig. 74 and Fig. 75) is also very comparable to Biber's (1988) Dimension 2 and will therefore also be labelled: 'Narrative vs. Non-narrative'. Although the normalisation bases are very different, the highest loading features are the same: past tense verbs (VBD) and third-person referents (TPP3). The perfect aspect (PEAS), which is the third highest-loading feature on Biber's Dimension 2, ranks eighth on PC2 (see Table 69), whilst public verbs (e.g., EXPLAIN, SAY, TELL), the fourth most important positive contributor to Biber's Dimension 2, corresponds broadly to the high contribution of verbs of communication (COMM) on PC2. However, unlike Biber's (1988) Dimension 2, the 'Narrative vs. Non-narrative' dimension to emerge from the present MDA involves a number of significantly loading features with negative contributions to the dimension. These include noun compounds (NCOMP), BE as a main verb (BEMA) and the modal *can* (MDCA).

The three ellipses of the TEC registers are noticeably "shifted" towards the middle of the biplot depicting texts on the first and second dimensions (Fig. 74). In the following sections, the reason for this "shift" towards the middle of the biplot will be explored. To this end, both the full table of feature loadings (Table 69) and graphs of variables such as Fig. 75 are examined to understand the linguistic specificities of these three textbook registers. Linear mixed-effects models were also computed for each principal component but, for reasons of space, only the most salient findings are reported in this chapter. All the models, tables and plots that were explored as part of the following analyses can be found in the Online Appendix 7.2.

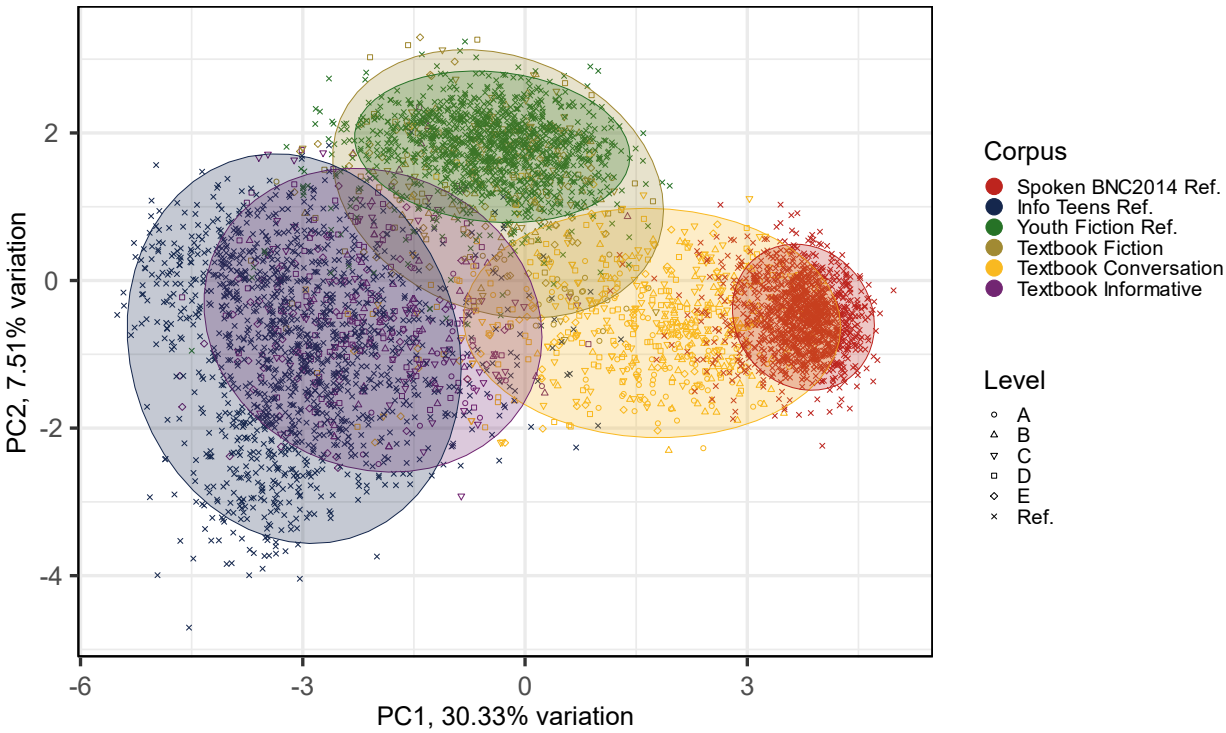


Fig. 74: Projection of the texts of the three subcorpora of the TEC and the reference corpora on PC1 and PC2

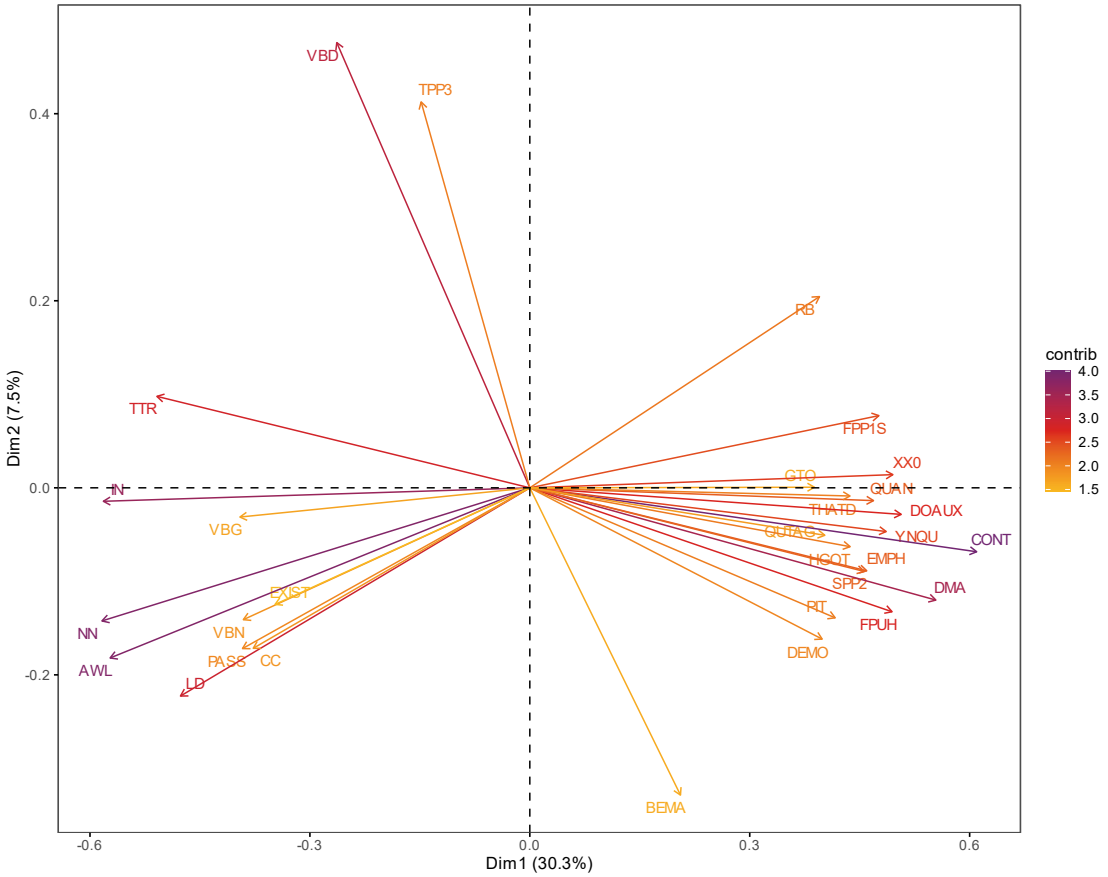


Fig. 75: Graph of the features with the strongest contributions to the first and second dimensions

Table 69: List of feature loadings (eigenvectors) in the Textbook English vs. 'real-life' English PCA-based model

	PC1	PC2	PC3	PC4
ACT	-0.10	0.01	-0.01	0.12
AMP	0.00	-0.05	-0.05	0.09
ASPECT	-0.08	0.10	-0.01	0.00
AWL	-0.21	-0.13	-0.06	-0.01
BEMA	0.08	-0.24	0.06	-0.02
CAUSE	-0.08	-0.12	0.06	0.14
CC	-0.14	-0.13	-0.09	-0.09
COMM	-0.03	0.19	0.03	0.20
CONC	-0.03	-0.03	-0.18	-0.06
COND	0.08	-0.02	-0.17	0.23
CONT	0.22	-0.05	0.03	0.00
CUZ	0.10	-0.14	-0.20	-0.18
DEMO	0.15	-0.12	-0.08	-0.04
DMA	0.20	-0.09	-0.04	-0.16
DOAUX	0.18	-0.02	0.06	0.00
DT	0.08	0.16	-0.24	-0.03
DWNT	-0.04	0.10	-0.12	0.09
ELAB	-0.07	-0.17	-0.04	0.07
EMPH	0.17	-0.07	-0.09	-0.03
EX	0.06	-0.02	-0.04	0.00
EXIST	-0.13	-0.09	-0.05	0.00
FPP1P	0.08	-0.02	0.09	0.19
FPP1S	0.17	0.06	0.06	0.08
FPUH	0.18	-0.10	-0.05	-0.19
FQTI	-0.07	0.03	0.01	0.14
GTO	0.14	0.00	-0.04	0.00
HDG	0.10	-0.07	-0.14	-0.12
HGOT	0.16	-0.05	-0.01	-0.11
IN	-0.21	-0.01	-0.08	0.01
JJAT	-0.05	-0.13	-0.25	0.07
JJPR	-0.03	-0.17	-0.03	0.21
LD	-0.17	-0.16	0.13	-0.04
MDCA	0.05	-0.21	0.11	0.22
MDCO	0.00	0.19	-0.15	0.10
MDMM	-0.02	-0.10	-0.14	0.17
MDNE	0.06	-0.02	-0.06	0.22
MDWO	0.07	0.11	-0.18	0.04
MDWS	0.06	-0.02	-0.01	0.25
MENTAL	0.11	-0.02	-0.05	0.16
NCOMP	0.00	-0.27	-0.05	0.03
NN	-0.21	-0.10	0.10	-0.07
OCCUR	-0.13	-0.02	-0.05	-0.06

PASS	-0.14	-0.13	-0.09	-0.10
PEAS	-0.06	0.12	-0.19	0.12
PIT	0.15	-0.10	-0.15	-0.07
PLACE	0.02	0.09	0.09	0.07
POLITE	0.09	0.00	0.20	0.11
POS	0.02	0.09	0.04	-0.05
PROG	0.09	0.08	-0.04	0.15
QUAN	0.17	-0.01	-0.16	0.01
QUPR	0.08	0.11	-0.12	0.21
QUTAG	0.15	-0.04	-0.07	-0.15
RB	0.14	0.15	-0.18	0.07
RP	-0.01	0.22	-0.09	0.15
SPLIT	-0.03	-0.11	-0.21	0.08
SPP2	0.17	-0.07	0.10	0.16
STPR	0.10	0.01	0.01	-0.04
THATD	0.16	-0.01	-0.14	-0.02
THRC	-0.05	-0.17	-0.15	-0.02
THSC	-0.02	-0.08	-0.27	0.07
TPP3	-0.05	0.30	-0.04	-0.15
TTR	-0.19	0.07	-0.02	0.16
VBD	-0.10	0.35	-0.05	-0.20
VBG	-0.14	-0.02	-0.14	0.12
VBN	-0.14	-0.10	-0.08	-0.07
VIMP	0.01	-0.07	0.21	0.21
WHQU	0.13	-0.02	0.20	0.07
WHSC	-0.09	-0.10	-0.20	0.05
XX0	0.18	0.01	-0.06	0.06
YNQU	0.18	-0.03	0.14	-0.02

Given that the first two dimensions appear to be functionally and linguistically analogous to Biber's (1988) dimensions, we can expect that many of the similarities and differences between Textbook English registers and situationally comparable, reference registers observed in Chapter 6 using additive MDA will be confirmed in the present analysis.

The third dimension (PC3), however, is not represented in Biber's (1988) model. Indeed, Fig. 76 shows that this third dimension appears to directly model some specificities of Textbook English as opposed to non-textbook, naturally occurring English. It is therefore labelled 'Pedagogically adapted vs. Natural'. Whilst the ellipses of the three reference corpora are entirely superimposed onto each other along PC3, with centroids around zero on the PC3 axis, the ellipses of the three TEC registers are notably shifted towards the positive end of the dimension. In addition, the elongated shapes of the ellipses of the textbook registers on this biplot shows that these texts cover a wide range of PC3 scores. Fig. 77 demonstrates that much of this

intra-register variation is driven by the proficiency levels of the textbooks: it displays the texts on PC3 and PC4 in exactly the same position as in Fig. 76 but, on Fig. 77, the ellipses correspond to the Level variable rather than, as in Fig. 76, the (Sub-)corpus variable. These ellipses show that, on average, level A textbook texts score highest on this third dimension and that, as the proficiency level of the textbooks increases, PC3 scores decrease.

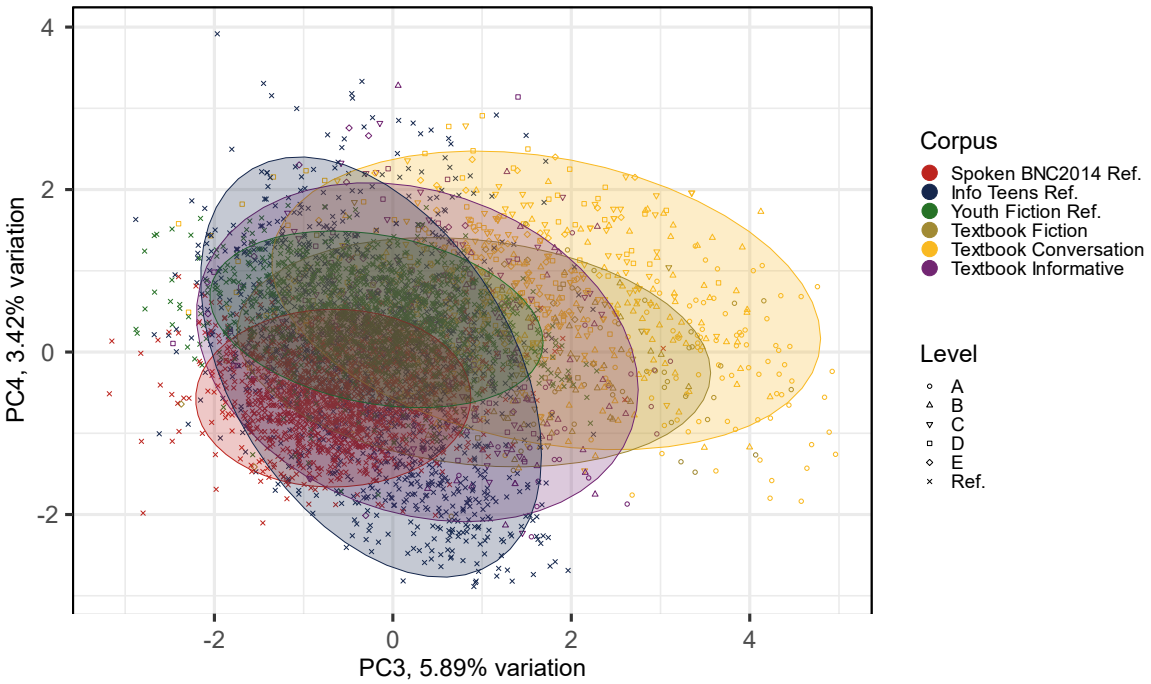


Fig. 76: Projection of the texts of the three subcorpora of the TEC and the reference corpora on PC3 and PC4

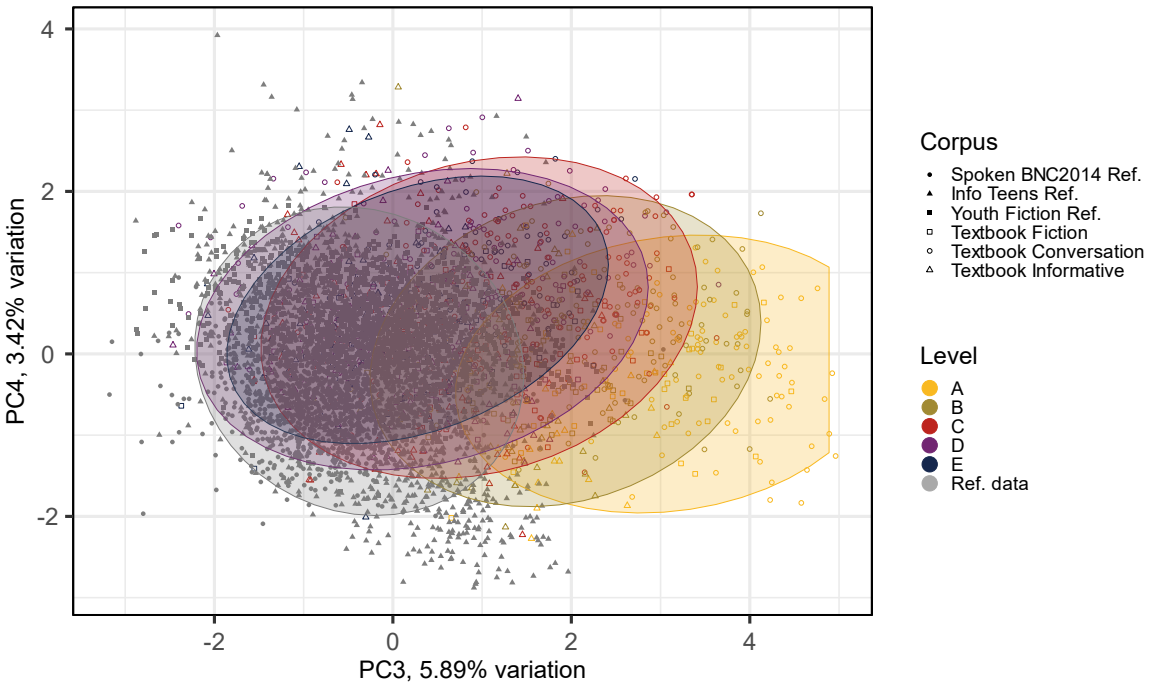


Fig. 77: Projection of the texts of the three subcorpora of the TEC and the reference corpora on PC3 and PC4 with ellipses representing the five textbook proficiency levels vs. the reference corpora

This effect is confirmed in the summary of the linear mixed-effects model computed to predict scores on this third 'Pedagogically adapted vs. Natural' dimension (see Table 70). Note that, in this and all other models explored in the following sections, the reference level is the Reference (Ref.) Conversation data, i.e., the Spoken BNC2014. As opposed to the PCA-based model of intra-textbook variation discussed in 7.3.1, only three registers are now modelled: Conversation, Fiction, and Informative. In addition, the Level variable now includes a sixth level for the Ref. corpora alongside the familiar proficiency levels of the TEC (A to E). This means that, on this third dimension, Table 70 shows us that Textbook Conversation texts from beginner (level A) textbooks score, on average, 4.25 points more than the intercept of -0.64, i.e., 3.61. Hence, we observe a particularly large difference between beginner Textbook Conversation texts and the transcripts of the Spoken BNC2014. Following the same logic, Table 70 shows that Textbook Fiction texts from the most advanced textbooks in the TEC have a mean PC3 score of -0.05 (= -0.64 + 1.29 + 0.43 + -1.03), which, by contrast, is remarkably close to the Youth Fiction mean PC3 score of -0.21 (= -0.64 + 0.43), which suggests that, on this dimension, there are probably no meaningful linguistic differences between these texts. For ease of interpretability, PC3 scores as predicted by the model are plotted on Fig. 78.

Table 70: Summary of the model: lmer(PC3 ~ 1 + Level + Register + Level*Register + (1|Source))

Predictors	Estimates	95% CI	*p*-value
(Intercept) [Conversation] [Ref.]	-0.64	-1.99 – 0.71	0.354
Level [A]	4.25	2.82 – 5.69	**<0.001**
Level [B]	3.09	1.66 – 4.52	**<0.001**
Level [C]	2.12	0.69 – 3.55	**0.004**
Level [D]	1.64	0.21 – 3.07	**0.024**
Level [E]	1.29	-0.15 – 2.73	0.078
Register [Fiction]	0.43	-0.92 – 1.79	0.533
Register [Informative]	0.43	-0.96 – 1.83	0.544
Level [A] * Register [Fiction]	-1.5	-2.89 – -0.12	**0.033**
Level [B] * Register [Fiction]	-1.39	-2.76 – -0.01	**0.048**
Level [C] * Register [Fiction]	-1.34	-2.71 – 0.03	0.056
Level [D] * Register [Fiction]	-1.35	-2.72 – 0.03	0.055
Level [E] * Register [Fiction]	-1.03	-2.41 – 0.35	0.142
Level [A] * Register [Informative]	-1.92	-3.35 – -0.49	**0.008**
Level [B] * Register [Informative]	-1.45	-2.86 – -0.03	**0.045**
Level [C] * Register [Informative]	-1.36	-2.77 – 0.05	0.058
Level [D] * Register [Informative]	-1.43	-2.84 – -0.02	**0.047**
Level [E] * Register [Informative]	-1.53	-2.95 – -0.11	**0.034**
	Random Effects		
σ^2	0.52		

$\tau_{00\ Source}$	0.48
ICC	0.48
$N_{\ Source}$	325
Observations	4980
Marginal R^2 / *Conditional* R^2	0.425 / 0.700

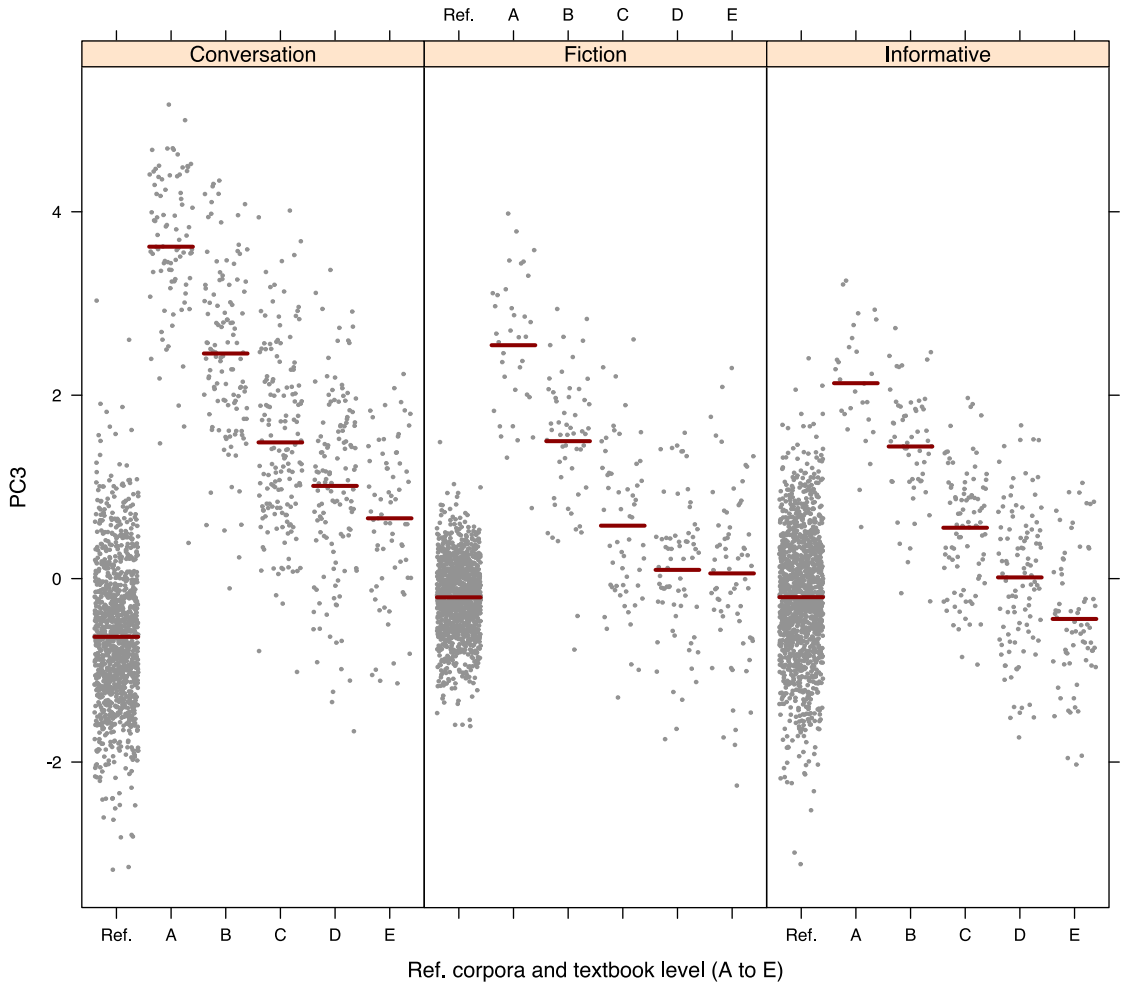


Fig. 78: Predicted PC3 scores of the texts of the TEC and the reference corpora

At first glance, the clear proficiency level patterns displayed in Fig. 78 would suggest that the negative loading features on this third dimension are all advanced linguistic features that are typically not taught until after learners have been acquainted with more basic lexico-grammatical phenomena. Looking at Fig. 79 and Table 69, however, we find that the linguistic features that make the greatest negative contributions to PC3 scores include subordinate clauses (THSC and WHSC) and causative subordinators (CUZ) per 100 finite verb phrases, as well as the frequency of attributive adjectives (JJAT) and determiners (DT) per 100 nouns – which are all features that are introduced early in the curricula followed by the textbooks included in the TEC. Only split auxiliaries (SPLIT) and the perfect aspect (PEAS) further

down the list are unambiguous examples of features not usually introduced until the third year of secondary EFL instructions.

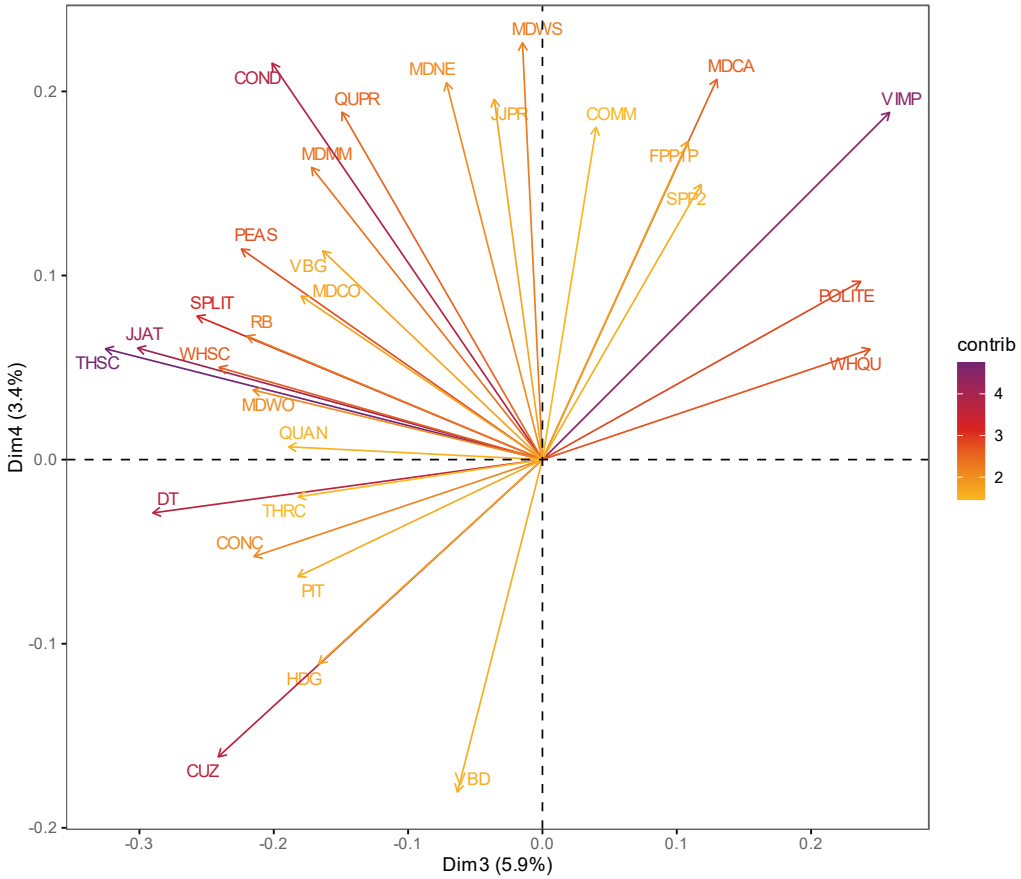


Fig. 79: Features that make the most important contributions to the third and fourth dimensions

Some of the features that characterise the positive end of the third dimension are reminiscent of the instructional end of the first dimension of the intra-textbook PCA-based model (see 7.3.1.1): e.g., imperative verbs (VIMP), WH-questions (WHQU), and second-person referents (SPP2). Indeed, many of the texts that score particularly high on the present third dimension model classroom dialogues. Such dialogues also include high frequencies of the modal verb *can* (MDCA), *yes/no* questions (YNQU), and politeness markers (POLITE) which also contribute to high PC3 scores, e.g.:

(231) **What**'s **your** name young man?
I'm Chad, Sir, Chad O'Malley.
Well I'm Mr Lloyd, the headmaster. **Remember** my name. Now, **act** your age and **stop** chatting!
Patrick, **can you** open the window, **please?**
Sorry Mrs Preston, I'm late. **Can** I come in**?**
It's alright for today Scarlett, but don't **forget** the room number next time. Well, children, let's start. Let's talk about kings and queens!
Can you give me the name of...the Queen of England**?**
Queen Elizabeth!
Well. That's easy. More difficult... **Can you** give me the name of a Norman king... Tom! **Can you** repeat the question**?**
... a Norman king!

William the Conqueror!
Congratulations Scarlett!
Wow!
And now [a] question for the champions. **Can you** tell me who this man is**?**
He had six wives. His first wife was Catherine of Aragon, his second wife was Ann Boleyn and...
Miss... Miss... I know! I know!
Yes Patrick?
It's King George.
No, it isn't.
Scarlett! **Can you** answer**?**
Henry VIII.
Brilliant! **Join** the school history club!
She's incredible! <TEC: Join the Team 6^{e}>

In general, this third dimension indicates that, as opposed to non-textbook language, Textbook English is characterised by high lexical density (LD) and, in particular, a high frequency of nouns (NN). Encouragingly, Fig. 78 shows that, on this 'Pedagogically adapted vs. Natural' dimension, texts from more advanced textbooks tend to score most like their corresponding reference corpora, although the gap between Level E textbook dialogues and the Spoken BNC2014 remains alarmingly wide.

The fourth dimension to emerge from this 'Textbook English vs. real-life English' PCA accounts for just 3.42% of the total variance. However, it is considered of relevance in the present study because it further contributes to identifying aspects of Textbook Conversation that distinguish this text variety from naturally occurring conversation. Whilst the ellipses of the other two TEC ellipses largely overlap with those of their corresponding reference corpora, it is clear from Fig. 76 that the vast majority of Textbook Conversation texts score higher on this dimension than those of the Spoken BNC2014. As shown on Fig. 79, the features that characterise its positive pole include conditional clauses (COND), the (semi-)modals *must, need* (MDNE), *will*, *shall* (MDWS), *may* and *might* (MDMM), the progressive (PROG) and perfect aspect (PEAS) and imperative verbs (VIMP), whilst its negative end is associated with causative subordinators (CUZ), hedges (HDG), and the past tense (VBD). The texts that score highest on this dimension are informative texts that provide advice (e.g., many texts from the teenkidsnews.com and teenvogue.com subcorpora of the Info Teens corpus) whilst those that score lowest are factual texts reporting on historical events (e.g., many of the texts of the factmonster.com, and historyforkids.net subcorpora of the Info Teens corpus). This could hint at a 'Factual vs. Speculative' dimension, were it not for the fact that hedging devices (HDG) are a major contributor to the 'factual' end of the dimension. Thus, this dimension does not lend itself well to a purely functional interpretation. In fact, it foremost differentiates between texts with simple verb forms and those with complex ones (modals, perfect and progressive aspects, etc.). The result is that a register that generally employs high rates of simple verb forms per FVP, e.g., natural face-to-face

conversation, scores low on this dimension whilst textbooks' synthetic representations of conversation feature many more complex verb forms and therefore score higher.

Given that this fourth dimension (at least partly) captures verb phrase complexity, it is not surprising that, as in the third dimension, it is considerably influenced by the different target proficiency levels of the textbooks. On this fourth dimension, the proficiency level effects are evidently driven by the fact that many of the aforementioned positive-loading features representing more complex verb forms are not introduced until the second or third year of English tuition. The mixed-effects models computed for PC4 scores make clear that textbook proficiency level affects dimension scores differently depending on the register under study. These effects are visualised in Fig. 80. Thus, we observe that, as the proficiency level of the textbooks increases, the dimension scores of the Fiction and Informative textbook subcorpora converge to that of their corresponding reference corpora whereas, curiously, the opposite pattern emerges for Textbook Conversation. The driving forces behind this effect will be examined in the following section (7.3.2.1), which explores the linguistic differences between the texts of the Textbook Conversation subcorpus and those of the Spoken BNC2014.

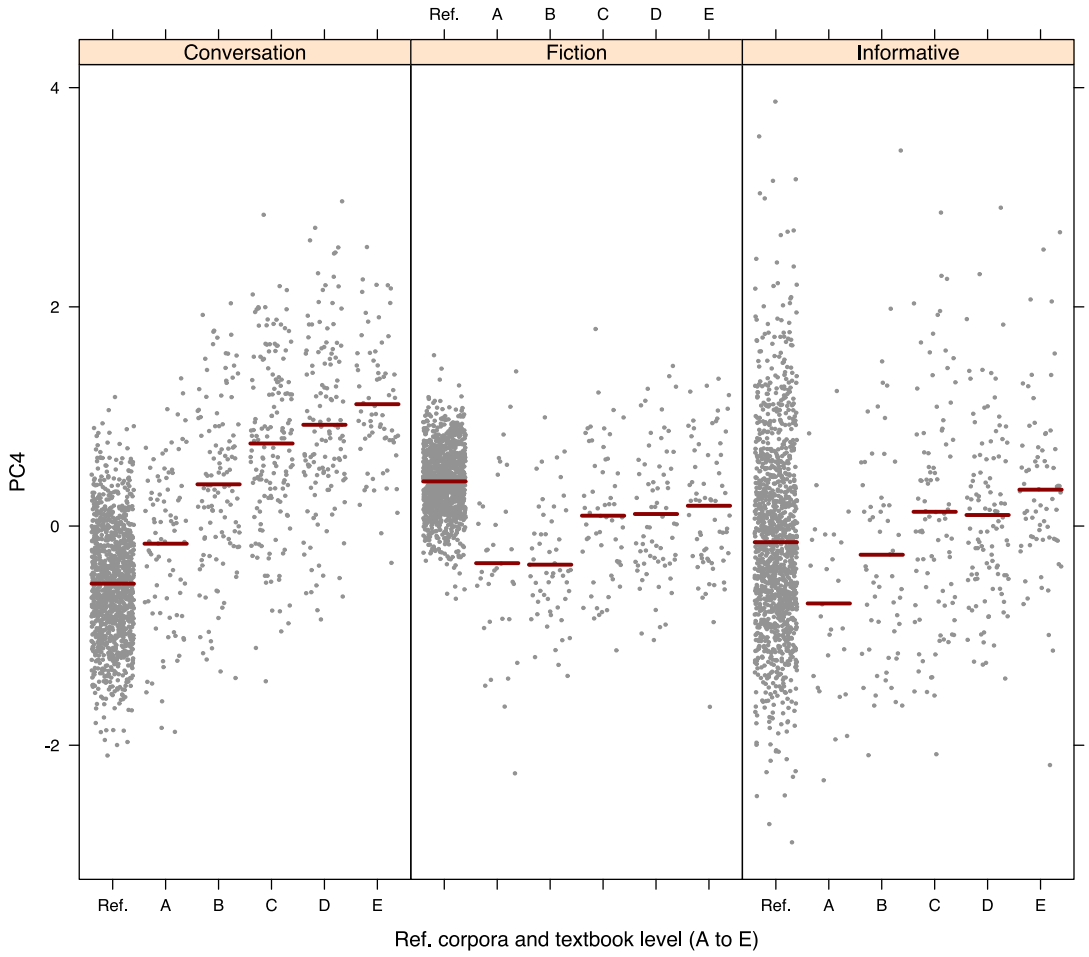


Fig. 80: Predicted PC4 scores of the texts of the TEC and the reference corpora

7.3.2.1 Textbook Conversation vs. the Spoken BNC2014

As seen in Fig. 81, the majority of Textbook Conversation texts differ quite substantially from the texts of the Spoken BNC2014 on the first 'Spontaneous interactional vs. Edited informational' dimension, which explains nearly a third of the total variance in the full data matrix. The projection of texts in Fig. 81 also shows that, whilst the reference conversation data forms a relatively tight cluster of texts (mean = 3.74, SD = 0.49), the ellipse of the conversation subcorpus of the TEC (mean = 1.56, SD = 1.25) spreads across a much larger area. On this first dimension, therefore, Textbook Conversation texts vary a lot more than the texts of the Spoken BNC2014 that resemble each other much more.

To explore this variation further, PC1 scores were modelled by entering Register, Level and their interactions in a mixed-effects model. The full model (see summary in Table 71) explains 92% of variation in PC1 scores (conditional R^2). Removing the random slopes and intercepts accounting for different text sources (see 7.2.9) does not lead to a substantial drop in predictive power (adjusted R^2 = 87%). Table 71 confirms that Textbook Conversation at all proficiency levels scores significantly lower on PC1 than the Spoken BNC2014 (here, the reference level). The results also show that the proficiency levels of the textbooks do interact significantly with scores on this first dimension but, as already observed on Biber's (1988) Dimension 1 in the additive MDA in 6.3.1.1, for the conversation register, not in the expected direction: the more advanced textbooks present dialogues that are even less similar to naturally occurring conversation than the beginner textbooks! This is illustrated on Fig. 81, which displays the predicted PC1 scores across the three reference corpora and their corresponding textbook registers subdivided by proficiency levels. It confirms that this unexpected finding is only true for the conversation register.

Table 71: Summary of the model: `lmer(PC1 ~ 1 + Level + Register + Level*Register + (Register|Source))`

Predictors	Estimates	95% CI	*p*-value
(Intercept) [Conversation] [Ref.]	3.71	2.46 – 4.96	**<0.001**
Level [A]	-1.73	-3.06 – -0.40	**0.011**
Level [B]	-1.65	-2.97 – -0.32	**0.015**
Level [C]	-2.08	-3.40 – -0.76	**0.002**
Level [D]	-2.47	-3.80 – -1.15	**<0.001**
Level [E]	-2.73	-4.06 – -1.40	**<0.001**
Register [Fiction]	-4.20	-5.45 – -2.95	**<0.001**
Register [Informative]	-6.75	-8.03 – -5.47	**<0.001**
Level [A] * Register [Fiction]	2.18	0.86 – 3.49	**0.001**
Level [B] * Register [Fiction]	1.71	0.41 – 3.01	**0.01**
Level [C] * Register [Fiction]	2.12	0.83 – 3.42	**0.001**
Level [D] * Register [Fiction]	1.93	0.64 – 3.23	**0.003**
Level [E] * Register [Fiction]	2.41	1.10 – 3.71	**<0.001**

Level [A] * Register [Informative]	3.37	2.01 – 4.73	**<0.001**
Level [B] * Register [Informative]	3.17	1.83 – 4.51	**<0.001**
Level [C] * Register [Informative]	3.24	1.90 – 4.57	**<0.001**
Level [D] * Register [Informative]	3.20	1.87 – 4.53	**<0.001**
Level [E] * Register [Informative]	3.14	1.80 – 4.48	**<0.001**
	Random Effects		
σ^2	0.59		
$\tau_{00\ Source}$	0.41		
$\tau_{11\ Source.RegisterFiction}$	0.12		
$\tau_{11\ Source.RegisterInformative}$	0.20		
ρ_{01}	-0.05		
	-0.48		
ICC	0.41		
$N_{\ Source}$	325		
Marginal R^2 / Conditional R^2	0.870 / 0.923		

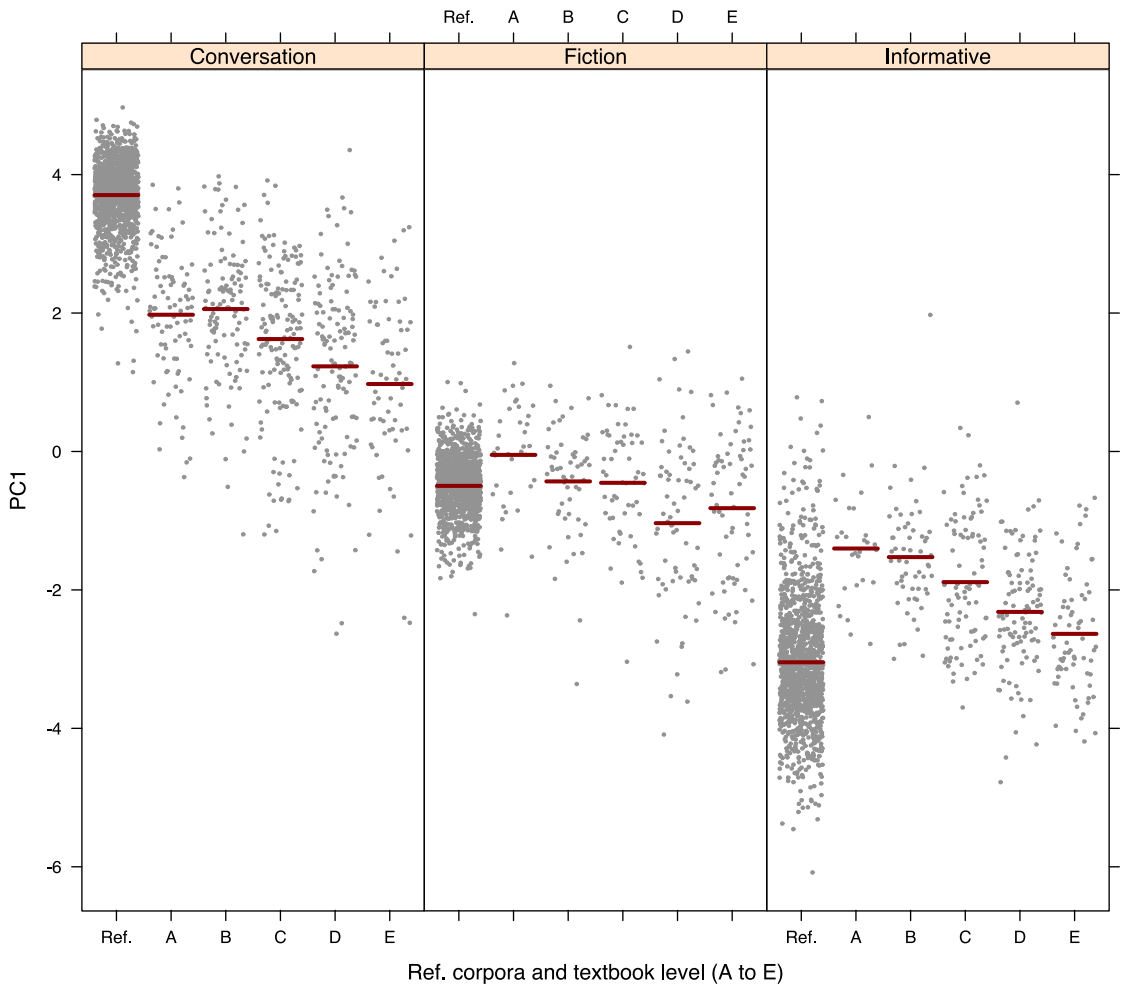


Fig. 81: Predicted PC1 scores of the texts of the TEC and the reference corpora

We have already noted that the first dimension to emerge from the present PCA-based MDA bears many similarities to Biber's (1988) Dimension 1. As shown in Fig. 74 and Table 69, many of the features with particularly high and low loadings are

shared. In particular, the upper end of the dimension is dominated by texts that contain high frequencies of many features of unplanned speech, e.g., discourse markers (DMA), fillers and interjections (FPUH) and causative subordinators (CUZ), as well as features typical of social interaction, e.g., questions and question tags (YNQU, WHQU, QUTAG), first-person and second-person references (FPP1S, FPP1P, SPP2), e.g., (232). All of these features are less frequent in Textbook Conversation than in the Spoken BNC2014, which is one of the reasons why, on average, the dialogues of the TEC score significantly lower on PC1 than the transcripts of the Spoken BNC2014, e.g., (233). As already observed in 6.4, this is likely due to the fact that, in their carefully crafted, scripted dialogues, textbook authors very rarely model any disfluencies in spoken interactions, e.g., in the form of hesitations, interruptions, backtracking, or repair. In addition, many Textbook Conversation texts are much more like mini-monologues than genuine interactions. On average, turns are much longer and feature far fewer signs of unplanned speech, e.g., (234).

(232) **yeah well** there's a few things **I** need to do in **um** town
mum was thinking of dropping **me** off at Zumba and then going
yeah well I need to pay **those** cheques in
can **you** pay **my** cheque in? and get some money out for the rent. can **you** pay **my** cheque in?
yeah
thanks
pay your cheque in and anything else **we** need in town?
I can't think but **we**'re gonna have to get up early
what**?** How early's early**?**
well I well not too early **we** just leave at nine **we** have to be ready and **I** have to have **my** breakfast before then so **I** can
okay
but **I**'m not too sure if **I**'m gonna go to Zumba or come to town with **you**
right
but either way **we**'re going to town at nine
okay what do **you** do in Zumba**?** What is **it?**
dance **it**'s just that **you** remember **those those** videos **you** used to have **like** aerobic dance videos
well I never had any but **yeah I** know what **you** mean
no no you've seen **me** do **them** before <BNC2014: SWU3>

(233) What are **you** up to at the weekend, Toby**?**
I'm going to go for a bike ride on Saturday. Do **you** fancy coming too**?**
I can't, **I**'m afraid. **I**'m going to help **my** dad with some gardening. **We**'re going to do some work for a neighbour.
That doesn't sound like the best way to spend **your** weekend. Gardening is hard work! And according to the forecast, the weather isn't going to be good.
I know. But the neighbour is going to pay **us** for **it**. And **my** dad's a gardener so he's got all the right tools.
Really? **I**'ll come and help **you**. **I mean**, if that's OK with **you** and **your** dad…
Sure. **We**'ll share the money with **you**: £10 an hour. But what about the bike ride**?**

I'll go on Sunday instead. The weather will probably be better then. Do **you** want to come?
Yes, please. **I** love bike rides. But let's go in the afternoon. **I**'ll be exhausted when **I** wake up! <TEC: Solutions Intermediate>

(234) **That's right**. This passion for having a perfect body is definitely a negative product of Hollywood. There seems to be an unwritten rule that in order to make it in the movies you must be beautiful. Your looks are almost like your business card and are often the only way to make a good first impression. But things get dangerous when they're taken to extremes. Examples of this are the body builders at Venice Beach, like in this picture, or people who spend tens of thousands of dollars a year on taking care of their looks and on plastic surgery - and they're not just celebrities. But is taking care of your body all bad**?**
As June said, it's when things are taken to extremes that they're harmful. There's also a very positive side to this Californian symbol. A lot of new kinds of sports were invented in California - here are some examples - and a lot of Californians lead very active and healthy lifestyles. In addition, California is often the country's leader when it comes to making public health laws. The Golden State was the first to ban smoking in workplaces, bars and restaurants in 1998 and now about half of the states in the US have similar laws. The state is also very active in limiting air pollution. But **what about** traffic in places like L.A.**?** <TEC: Green Line New 5>

Another factor that contributes to lower PC1 scores is the fact that some of the features with high loadings on PC1 are more likely to be used by language users who are physically in the same space, e.g., demonstratives (DEMO) and the pronoun *it* (PIT). Such features are equally frequent in the transcripts of the textbooks' video materials as in those of the Spoken BNC2014, but less frequent in audio-only materials and it is worth nothing that audio materials, which make up the largest proportion of the Conversation subcorpus of the TEC. By virtue of sharing a common environment, the speakers of the Spoken BNC2014 can also afford to resort to much vaguer language. This difference is also captured on this first dimension. In addition to the extensive use of demonstratives and the pronoun *it*, it is observable in the considerably more frequent use of quantifiers (QUAN), quantifying pronouns (QUPR) and hedges (HDG) than in Textbook Conversation.

(235) oh right so you **sort of** get so you **sort of** get things that **kind of** yeah I guess localised
yeah so you you got a potential for these vast metropolises filled with **loads of like** diverse em cultures and backgrounds **and stuff** but you end up getting **these** close communities em so I wanted to explore **that** idea and and in the video at the end when asked how what what would be the role of an artist em to to explore **this** idea or to confront **it** and em he suggested that perhaps the artist could expand em **sort of** enlarge the the garden gnome to the size of the statue of liberty **or something** just just poke fun of out of **it** and em so so that that was part of the reasoning for **this** first art project that I was going to do em and entering into of course what I thought I was going to be doing which was creative education they asked us to get completely obsessed with

something just one particular theme
oh right
just to explore what **it** was to to get obsessed and eh I'd recently bought Minecraft so I decided I'd go for **that** and just learn absolutely everything about **it** and em eh **something** that came across when I was looking at **bits and bobs** is that **a lot of** people on Minecraft servers are in their own little worlds they tend to make **these** monuments to Mario **it**'s a **It**'s a very easy statue to make there's there's pixel art of Mario and **it** happens **all** the time so I had **sort of** combined those ideas in my head and thought it would be **a bit of** fun to make a **sort of** combination of a garden gnome the statue of liberty and super Mario
oh wow
in eh on on the server and eh so when I announced **it** to people that **sort of** got excited about **it** and **sort of** made an island for **it** to sit on and eh **some** people were trying thought they might make a little **sort of** hidden illuminati style base
wow <BNC2014: SCJL>

Since the positive- and negative-loading features that contribute to each dimension are complementary, it is important to not only consider which positive high-loading features are absent from or less frequent in the Conversation subcorpus of the TEC to understand the nature of Textbook Conversation, but also which negative-loading ones are markedly more frequent in Conversation subcorpus of the TEC than in the reference Conversation corpus. These include nouns (NN), prepositions (IN), longer words (AWL), a higher lexical density (LD), and lexical diversity (TTR). As already noted in 6.4, this shows that, on average, Textbook Conversation is considerably more nominal than naturally occurring conversation. The fact that dialogues from more advanced textbooks score even lower than beginner textbooks on PC1 is due to higher z-scores of these features. Excerpts (227), (228) and (234), for instance, are examples of particularly low-scoring texts from advanced textbooks (level E), whereas (236) stems from a second-year secondary school textbook and, on this dimension, is situated in the same region as the majority of the Spoken BNC2014 texts. It features much shorter turns than the average textbook dialogue.

(236) **This** is lovely.
Isn't **it** lovely**?**
What is **it** exactly**?**
It's a coffee machine.
Oh, **yes**. **Of course**. Is **it** battery powered**?**
No, **it**'s mains powered.
Look, the cable's here, under the base. If **you** press **this** button, the plug appears.
That's clever. **I** love **it**. **It**'s perfect for my kitchen at home. **I**'ll come back later today and buy **it**.
Would **you** like to try a cup before **you** go**?**
I'm sorry?
A cup of coffee?
Oh, **no** thanks. **I** never drink coffee. Horrible stuff.
<TEC: Solutions Pre-intermediate>

As already explained in 7.3.2, the distribution of texts along the third dimension (PC3) confirms that, as opposed to real-life conversation, Textbook Conversation tends to feature exclusively well-formed polite interactions, with very few disfluencies. Whilst the dialogues of the more advanced textbooks of the TEC score closest to the reference Spoken BNC2014 texts (see Fig. 81), a noticeable distance can nevertheless be observed between the two varieties. On this third 'Pedagogically adapted vs. Natural' dimension, it is mostly driven by the high nominal content of Textbook Conversation leading to high noun counts (NN), high lexical density (LD), and proportionally fewer of those nouns being qualified by an attributive adjective (JJAT), as well as generally fewer causative subordinators (CUZ), subordinate clauses per finite verb phrases (THSC and WHSC) and general adverbs (RB). In addition, the modals *could* and *would* and phrasal verbs (as imperfectly captured by the RP variable) are, across all proficiency levels, less frequent in Textbook Conversation than in the Spoken BNC2014, e.g., (237). These three features also contribute to negative PC3 scores.

(237) I **would** like something to extract all the vegetable
I don't know I I like the idea of eating **very** little meat I **really** love molluscs and they can't and no one else even likes them why can't I have them?
I I doubt they're the worst I mean
do you think should I investigate **whether** the fishing practices **that** get molluscs are harmful to the environment?
yes I also think just reducing the amount you **would** eat **would** be doing a favour to the animals in some ways
but you should also look at erm **who** was involved with the farming industry problem okay? there might be […]
there might be people treated **badly**
mm mm but that's almost in everything and once you **really** get **down** to it **exactly**
like you you **basically**
it's terrifying
stop eating
so that's **probably** so that's why I was a bit like well I **couldn't** commit myself to this erm
no no
lab meat thing **cos** there'll be cruelty somewhere
yeah and
cos there's cruelty in everything there's **probably** cruelty in this potato.
cruel-free parsnips
and if you eat things that are **only** grown or made **locally** it's **very** expensive. mm mm
you know erm you know if if I had all the money in the world it **would** be very nice <BNC2014: SFC2>

On the fourth dimension, more advanced Textbook Conversation texts are situated further away from the reference Spoken BNC2014 than beginner and intermediate ones (see Fig. 80). One of the driving factors behind this surprising finding is certainly

that textbook authors use printed dialogues and audio and video materials to introduce new vocabulary and "recycle" previously introduced lexical items, leading to higher type/token ratios in more advanced textbooks, even though real-life conversations tend not to be characterised by particularly high lexical diversity, e.g., compare (237) and (238).

(238) **Good** morning Mr. Stone. **Good** morning. So, first, could you tell us more about the movie?
Sure, well... As a historian, **Modern** Times seems particularly **interesting** to me because Chaplin showed the effect of mechanization on the lives of people in the **modern industrialized** world. The film is about the hardship of an **ordinary** man who struggles to survive in the **depressed** economy of 1930s America. The factory scene is just **brilliant**! It is exactly what modernity was about at that time: the assembly line, the division of labour, mass production and the **daily** grind of many **industrial** workers. They were exploited by their bosses, who made them work from dawn to dusk in very **tough** conditions. The film shows the workers' **repetitive** labour and how some of them went **crazy** as they repeated the **same** task over and over again, tightening bolts for instance.
Some people say that industrialization has only had **negative** effects and they are **nostalgic** about the past, when things were **simple** and technology had not spoiled everything yet. Do you agree with that?
No, I don't quite agree with them because industrialization was also a progressive revolution; it did have very positive effects. […]
<TEC: Piece of Cake 4e>

7.3.2.2 Textbook Fiction vs. the Youth Fiction corpus

As already observed in the comparative additive MDA (see 6.3.2), of the three TEC registers compared to reference corpora, Textbook Fiction is most similar to its corresponding target reference corpus, Youth Fiction. Where differences are observed, they are mostly due to beginner textbooks having not yet introduced more advanced grammatical features. The strong proficiency level effects observed along the second 'Narrative vs. Non-narrative' dimension (see Table 72) are largely driven by the absence of the past tense in Level A textbooks (see also 6.3.2.2). Other high positive-loading features on PC2 that are almost entirely absent from Levels A and B textbooks include the modals *could* and *would* (MDCO and MDWO), and the perfect aspect (PEAS). In addition, the PC2 scores of Level A Textbook Fiction texts are lower due to a strong over-representation of two features with some of the largest negative-loading weights on this second dimension (see Table 69): the modal *can* (MDCA) (median = 6.25 per 100 FVPs) as compared to Youth Fiction (median = 1.65 per 100 FVPs) and BE as a main verb (BEMA) (19.2 per 100 FVPs compared to 14.2). It is, however, important to remember that there are relatively few Level A texts in the Fiction subcorpus of the TEC so that these figures ought to be interpreted with caution. Nonetheless, excerpt (239) constitutes a representative

example of a narrative text from a beginner EFL textbook: it relies on present tense narration and features high frequencies of BE as a main verb and the modal *can*.

(239) B) Holly **is** at home with her two guinea pigs, Mr Fluff and Honey. They live in the kitchen. But they **are**n't in the kitchen now. They'**re** in Holly's room. It'**s** fun for the guinea pigs on the floor. They **can** explore - everywhere in the room! C) Ding-dong! Who'**s** at the door? It'**s** Holly's best friend Olivia.
D) After Olivia's visit Holly **can** only see Honey under her bed. But where'**s** Mr Fluff? Holly **is**n't happy. She'**s** got a problem. There's an English test today, and she hasn't got her lucky charm with her in the classroom. Where **is** it? Maybe it'**s** in her schoolbag. Oh, but here'**s** the tutor, Mr Swindon, with Luke, Dave - and a new boy in the tutor group.
A) It'**s** a week before project day. The students in Mr Swindon's tutor group have got lots to do. The students **can** work in groups, in pairs, or alone. Who has got the best audio presentation? Dave and Luke **can** do the recordings for everyone in the TTS recording studio.
<TEC: Green Line New 1>

Table 72: Summary of the model: lmer(PC2 ~ 1 + Level + Register + Level*Register + (1|Source))

Predictors	Estimates	95% CI	*p*-value
(Intercept) [Conversation] [Ref.]	-0.52	-1.27 – 0.24	0.183
Level [A]	-0.54	-1.35 – 0.28	0.195
Level [B]	-0.15	-0.96 – 0.66	0.721
Level [C]	0.08	-0.72 – 0.89	0.842
Level [D]	0.04	-0.76 – 0.85	0.915
Level [E]	0.07	-0.75 – 0.89	0.866
Register [Fiction]	2.27	1.51 – 3.03	**<0.001**
Register [Informative]	-0.46	-1.24 – 0.32	0.25
Level [A] * Register [Fiction]	-1.01	-1.82 – -0.21	**0.014**
Level [B] * Register [Fiction]	-0.25	-1.04 – 0.54	0.532
Level [C] * Register [Fiction]	-0.21	-1.00 – 0.58	0.602
Level [D] * Register [Fiction]	-0.41	-1.20 – 0.38	0.307
Level [E] * Register [Fiction]	-0.47	-1.26 – 0.32	0.246
Level [A] * Register [Informative]	0.49	-0.34 – 1.33	0.246
Level [B] * Register [Informative]	0.59	-0.22 – 1.40	0.154
Level [C] * Register [Informative]	0.26	-0.54 – 1.06	0.524
Level [D] * Register [Informative]	0.55	-0.26 – 1.35	0.183
Level [E] * Register [Informative]	0.33	-0.49 – 1.14	0.431
	Random Effects		
σ^2	0.45		
$\tau_{00\ Source}$	0.15		
ICC	0.25		
$N_{\ Source}$	325		
Observations	4980		
Marginal R^2 / Conditional R^2	0.671 / 0.753		

Textbook Fiction texts that score within the range of the Youth Fiction corpus on the second dimension are characterised by high frequencies of past tense (VBD) and perfect aspect (PEAS) verbs, third-person referents (TPP3) and verbs of communication (COMM), but also the modal *could* (MDCO), phrasal verb particles (RP) and adverbs (RB). Note, however, that the latter two features are key contributors to the slightly lower average scores of intermediate and advanced Textbook Fiction texts compared to the Youth Fiction (see Table 72 and Fig. 82). Counter to the author's expectations, lexical diversity (TTR) or average word length (AWL) do not contribute to the observed minor differences between the more advanced Textbook Fiction texts and the reference Youth Fiction corpus.

(240) 'There's no need to be in such a rush,' **said** Jem, so **close** now that Joe **could** see **his** broken teeth. 'We got a lot of catching **up** to do.'
'Yoo-hoo! Joe!'
The voice **came** out of nowhere, making **them** all jump. It **was** a girl's voice, and when **he** looked **up**, Joe **saw** that Lil **was** hustling down the street towards **them**. **Her** cheeks **were** flushed with excitement: it **had** been **her** first night in **her** show at the theatre, **he realised**, and now **she was** on **her** way to the party. **She was** wearing a hat wreathed in poppies and **had** a crimson scarf at **her** neck that **deepened** the red of **her** lips and cheeks, and made **her** dark hair look glossier than ever.
'I didn't expect to see you **out** here!' **she said** as **she** bounced **up** to him. **Then she** took in the three men. 'Good evening,' **she said** to **them**, **brightly**. <Youth Fiction: Woodfine 2015: The Mystery of the Clockwork Sparrow>

(241) **He ran back** to Bill. On the way **he picked up** a stick. As **he came** over the hill, **he ran** at the boar, hitting it again and again. It **turned** to face Colm. There **was** blood on its tusks. Bill **tried** to pull himself away, leaving a trail of blood behind **him**. Colm raised the stick high and brought it **down** on the boar's head. The boar **snorted**, but instead of running at Colm, it **turned back** to Bill. Colm **threw** the stick **down** and **grabbed** Bill's gun. **He was** shaking as **he raised** the gun to **his** shoulder. Colm **knew** that if **he shot** the animal in the back, it would only make it wild. **He let** out a scream, a long, loud scream. The boar **turned round**. For a moment it **stared** at **him** with its small black eyes. Then it **lowered** its head and **ran** towards **him**. <TEC: Access G 5>

Fig. 82 clearly shows that, aside from Level A Textbook Fiction, the narrative texts of the TEC are, on average, largely comparable to those of the Youth Fiction on this second dimension. Encouragingly, the Register*Level effect plots in 7.3.2 show that this is also true on the third and fourth dimensions (see Fig. 78 and Fig. 80). It is, however, worth noting that there is a great deal more variation within the Textbook Fiction subcorpus than there is across the much larger Youth Fiction corpus, as evident from the large range of predicted dimension scores on all dimensions and across all proficiency levels (see, e.g., Fig. 74). The Textbook Fiction subcorpus being relatively small (285 texts, ca. 241,500 words) and narrative texts being rather rare in the three French textbook series (see 3.3.1.4), further data from additional

textbooks would be needed to confirm the trends concerning Textbook Fiction reported in this and the previous chapter.

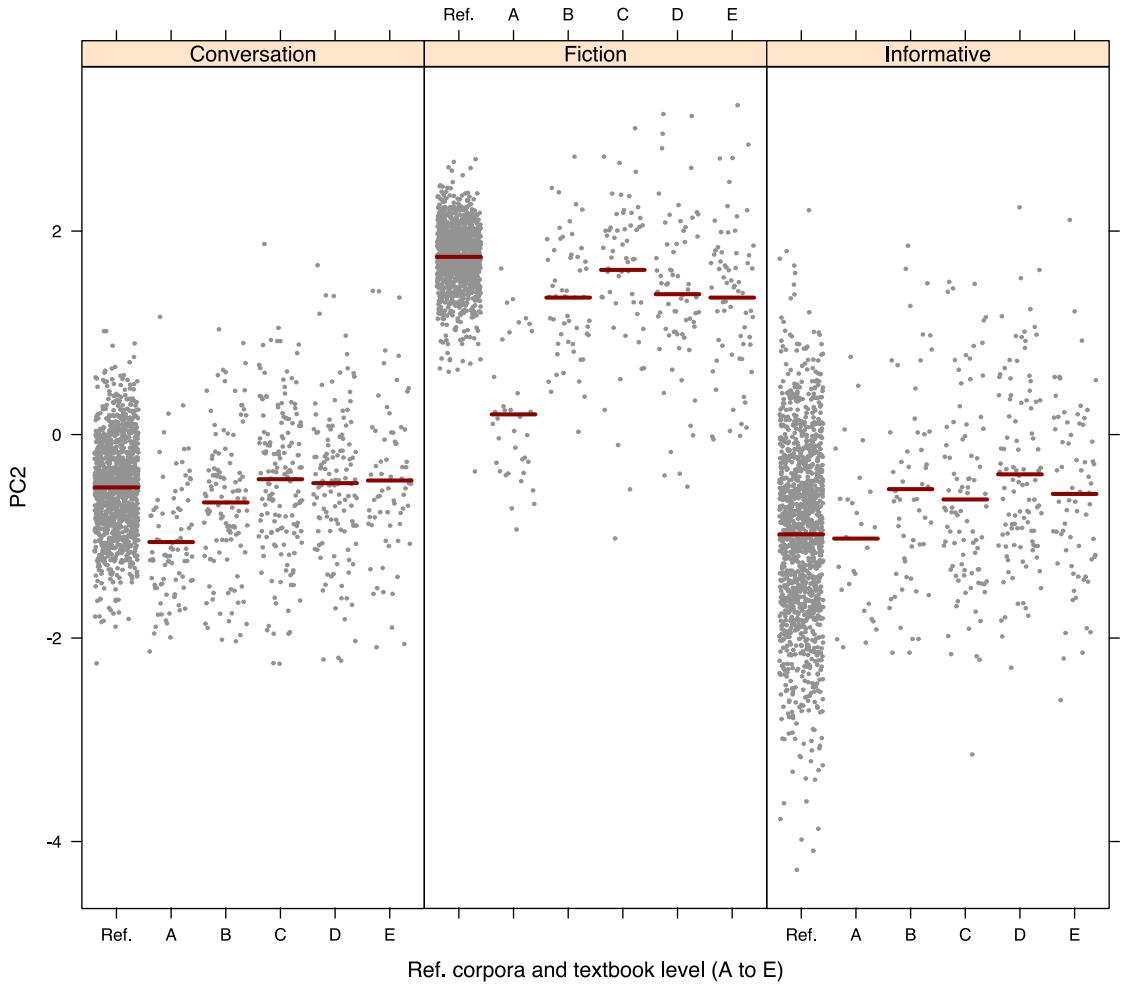


Fig. 82: Predicted PC2 scores of the texts of the TEC and the reference corpora

7.3.2.3 Textbook Informative vs. Info Teens

The most striking differences between the Textbook Informative subcorpus and Info Teens can be observed on the first dimension. Echoing the results of Chapter 6, Fig. 73 and Fig. 74 indicate that some of the Textbook Informative texts are closer to the interactional, "oral-like" end of the dimension than the Info Teens reference corpus. Fig. 74 also shows some overlap between the ellipses of the Textbook Informative and the Textbook Fiction subcorpora. Thus, we find that a proportion of textbook texts blur the otherwise well-defined register-based distinctions between the three reference corpora that emerge from the combination of the first and second dimension on Fig. 74. As confirmed by the model summarised in Table 71 and visualised in Fig. 81, however, many of the Textbook Informative texts that are not within the ellipse of the Info Teens corpus stem from beginner textbooks. In fact, on the first, third and fourth dimensions, the more advanced informative texts are much more similar to the corresponding reference corpus than the less advanced ones. These proficiency level patterns have already been explored in 6.4 and 7.3.1. Interestingly, the differences

observed between the texts of the beginner Textbook Informative subcorpus and those of the Info Teens corpus are only partially due to more complex linguistic features not being introduced until after the first few years of EFL tuition. Fig. 83 displays normalised counts for a sample of features with high absolute loadings on PC1 that most contribute to the differences observed between textbook and non-textbook informative texts.

Three groups of features can be identified on Fig. 83. The first group consists of complex grammatical features such as passive constructions (PASS) and non-finite present participial constructions (VBN) which, as expected, are very infrequent in beginner informative texts but, as Fig. 83 makes clear, whose frequencies progressively increase as learners progress. Second, we find features such as average word length (AWL) and existential verbs (EXIST) which also have considerably lower rates in beginner textbooks than in more advanced ones, but whose rates in Level E textbooks do not, on average, reach rates quite as high as those observed in the Info Teens corpus. Finally, the third group of interest consists of features that are, across all proficiency levels, more frequent in the Textbook Informative subcorpus than in the reference Info Teens. As these features have high positive-loading weights on PC1, they contribute to the observed shift of Textbook Informative texts towards the middle of the plots in Fig. 72 and Fig. 74. They include singular first- and second-person referents (FPP1S and SPP2) and verbal contractions (CONT) and negation (XX0), which are, on average, considerably rarer in the texts of the Info Teens corpus than in Textbook Informative writing. These differences are exemplified in (242) and (243). The negative-loading features that are, on average, rarer in Textbook Informative texts yet typical of the Info Teens are highlighted in (242), whilst the words in bold in (243) correspond to positive-loading features that contribute to Textbook Informative texts scoring higher on the first dimension than the Info Teens.

(242) Tennyson was the son of an intelligent but unstable clergyman in Lincolnshire. His early literary attempts **included** a play, The Devil and the Lady, **composed** at 14, and poems **written** with his brothers Frederick and Charles but **entitled** Poems by Two Brothers (1827). In his three years at Cambridge, Tennyson wrote a prize-winning poem, Timbuctoo (1829), and Poems, Chiefly Lyrical (1830) and began his close friendship with Arthur Henry Hallam, son of the historian Henry Hallam. Upon the death of his father in 1831, Tennyson became responsible for the family and its precarious finances. His volume Poems (1832) **included** some of his most famous pieces, such as "The Lotus-Eaters," "A Dream of Fair Women," and "The Lady of Shalott." In 1833 he **was overwhelmed** by the sudden death of Hallam. […] Tennyson passed his last years in comfort. In 1883 he **was created** a peer and occupied a seat in the House of Lords. Throughout much of his life he was a popular as well as critical success and **was venerated** by the general public. Ignored early in the 20th century, Tennyson has since **been recognized** as a great poet, notable for his mastery of technique, his superb use of sensuous language, and his profundity of thought. <Info Teens: factmonster.com>

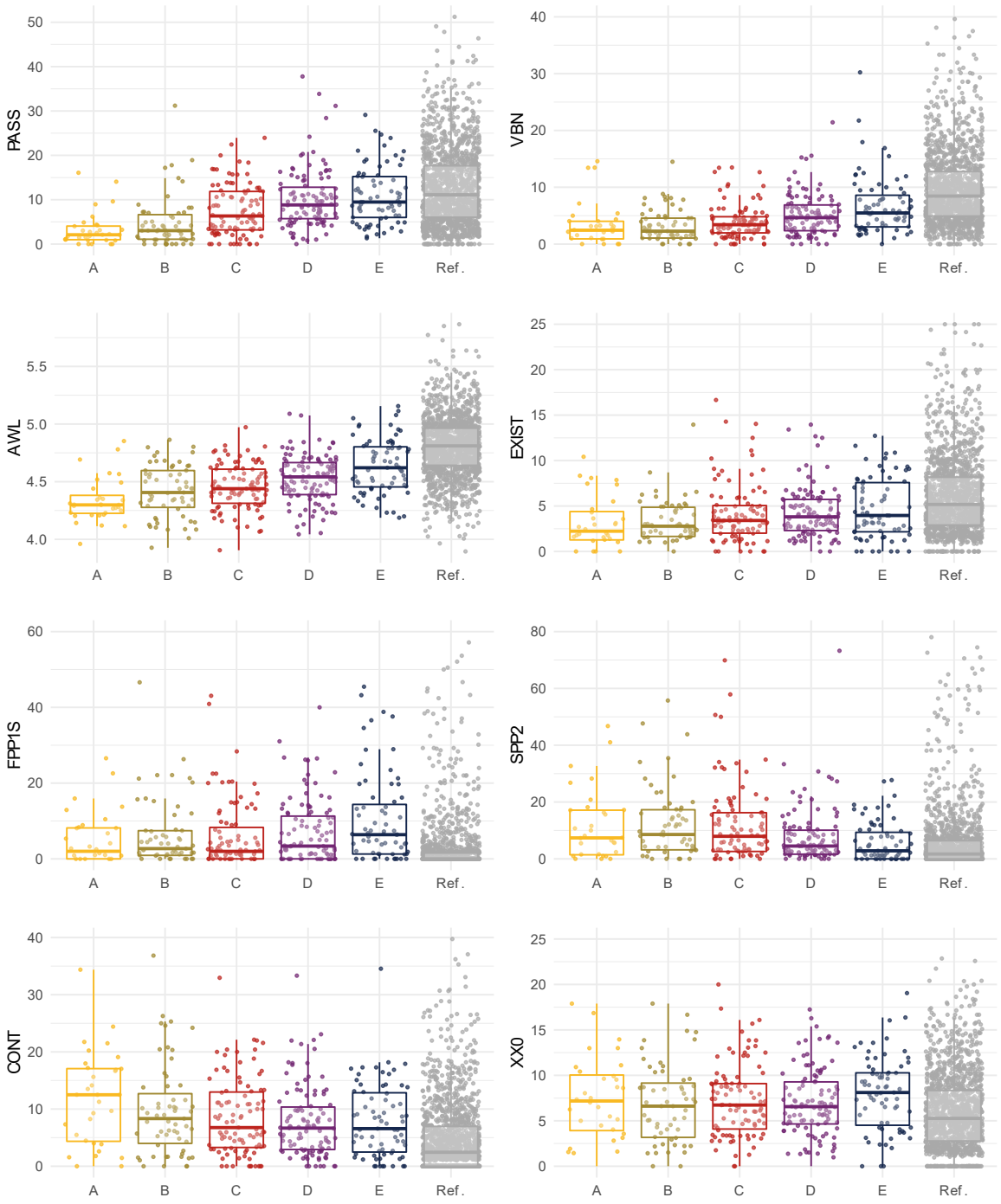


Fig. 83: Normalised counts of selected features with salient loadings on PC1 in the Textbook Informative subcorpus (Levels A to E) and the reference Info Teens corpus (Ref.)

(243) **You** might be surprised at the number of rather unusual sports that exist around the world. Mostly, they are little known outside the areas where they were invented – though occasionally they have gained international recognition. Here are some examples – but, if **you're** interested, have a look on the web. **You** may find other, even crazier, ones!
Sandboarding
Of course there **can't** be many people who **don't** know what snowboarding is, but how about sandboarding? The basic principle behind the two sports is the same; start at the top of a slope and use a board to get **you** to the bottom. But whereas snowboarding is practised on freezing

> cold snowy mountain tops, sandboarding takes place on sand dunes by sunny beaches or in the desert. It**'s** popular in many countries, including Australia, Namibia and South Africa. The quickest way of getting to the bottom involves standing with both feet on a board and weaving from side to side while trying not to fall off. If this sounds a little bit adventurous **you** could always just get on **your** stomach and slide down. Either way, it**'s** a lot of fun! However, **don't** forget to keep **your** mouth closed. <TEC: English in Mind 4>

The less pronounced differences observed between the texts of the Info Teens and those of the Textbook Informative subcorpus on the second 'Narrative vs. Non-narrative' dimension (see Fig. 82) are largely driven by past tense verbs: we know that they are largely absent from Level A textbooks, but, as soon as this tense and, to a lesser extent, the perfect aspect are introduced, textbooks from Level B onwards tend to feature more past and perfect aspect verbs in informative texts than are generally found in the Info Teens texts. This explains the small overlap between the Textbook Informative ellipse (purple) and the ellipses of the fiction corpora (light and dark green) on Fig. 74.

7.4 Discussion and limitations

In sum, the results of the present chapter can be said to refine the trends observed in Chapter 6. Many of the characteristics of the PCA-based multi-dimensional model of intra-textbook variation presented in 7.3.1 coincide with patterns observed in the intra-textbook additive MDA carried out in 6.3.1. The results of the second additive MDA (6.3.2), which mapped the texts of the Conversation, Fiction and Informative subcorpora of the TEC and of three matching reference corpora onto Biber's (1988) dimensions of general English (6.1.1), are also consistent with the new PCA-based multi-dimensional model of Textbook English proposed in 7.3.2.

In answer to the first set of research questions (see 7.1), considerable register-based variation can be observed in the language of school EFL textbooks. In fact, a much larger proportion of intra-textbook variation can be attributed to register differences than to proficiency levels, country of use, or of any potential idiosyncrasies of specific groups of authors or editorial policies (as captured by the textbook series variable). In particular, the visualisations of the model of intra-textbook variation revealed distinct clusters for the texts of the conversation, instructional and informative registers on the first few dimensions (as shown, e.g., in Fig. 56 and Fig. 57). The linguistic interpretation of Table 68, which lists the loadings of the 61 features entered in the model of intra-textbook variation, highlighted the defining characteristics of these different text registers frequently featured in secondary school EFL textbooks. The few proficiency level effects that emerged from the model were explained and illustrated with textbook excerpts. Additionally, significant interactions between register-based and proficiency-level-based variation – most notable in the fiction register (as already observed in Chapter 6) – were also examined. No notable

differences were observed across the three different countries of use of the textbooks (more on this below), nor were significant patterns of linguistic variation associated with any specific textbook series.

The second set of research questions asked to what extent Textbook English registers differ from situationally similar, naturally occurring registers and whether there might be pedagogically relevant differences between different textbook series and/or the proficiency level of individual textbook volumes. These questions were addressed in the second PCA-based multi-dimensional model of Textbook English (7.3.2). The most striking differences were observed in the textbooks' portrayal of spoken, conversational English. Section 7.3.2.1 examined in some depth the constellation of features that make most textbook dialogues fundamentally different from naturally occurring conversation (as captured by the Spoken BNC2014). These include low frequencies of many features that are typical of spontaneous speech, including fillers and interjections, discourse markers and causative subordinators, as well as markers of interactional discourse such as tag questions, WH-questions, pronouns, as well as hedges and demonstratives in communicative situations in which interlocutors share a common environment and, in many cases, common knowledge.

In fairness to the authors and editors of the textbooks featured in the TEC, some of the observed differences may be due to what is arguably an overly simplistic textbook register classification scheme (see 3.3.1.4). Indeed, as shown in some of the excerpts featured in this and previous chapters, the Conversation subcorpus of the TEC also includes radio and TV interviews which are situationally quite different to the informal everyday conversations among friends and relatives that constitute most of the texts of the Spoken BNC2014. Hence, future studies ought to consider re-annotating the texts of the Conversation subcorpus of the TEC to separate texts designed to emulate private, casual conversation from those that attempt to replicate public and broadcasted discussions. This second subset of texts could then be compared to a corpus of TV and radio language. Though this is likely to somewhat reduce the observed gap between how spoken English is portrayed in EFL textbooks and how it is spoken outside the EFL classroom, the trends described in the present and previous chapters can nonetheless be expected to remain largely the same. This prediction is motivated by the results of a study preliminary to Chapter 6 which was conducted before the Spoken BNC2014 was available and therefore relied on a corpus of TV captions and subtitles instead (a subset of the BBC corpus used in Fankhauser forthcoming). This study only included the German and French subcorpora of the TEC but it showed an equally wide gap between Textbook Conversation and (pseudo-)spoken language as observed in TV news, shows and series (Le Foll 2017).
Some noteworthy differences were also observed between the texts of the Informative subcorpus of the TEC and those of the Info Teens corpus (7.3.2.3). As compared to the Conversation register, interactions with the Informative subcorpus of the TEC

and the proficiency level of each textbook texts had a much stronger mediating effect on the placement of Textbook Informative texts on several dimensions of variation. Reassuringly, most of the informative texts featured in the more advanced texts closely resemble those of the reference corpus on all relevant dimensions of variation.

Section 7.3.2.2 confirmed the findings of previous chapters by demonstrating that, excluding the texts from beginner textbooks, Textbook Fiction is closest to its corresponding reference corpus on all dimensions of variation examined. In many respects, this finding is easily explained: unlike dialogues and informative texts which are mostly crafted especially for pedagogical purposes, many of the narrative texts featured in school EFL textbooks are extracts of 'authentic', published, popular novels and short stories – samples of which may even feature in the Youth Fiction corpus. That said, a couple of caveats also deserve mention. First, Textbook Fiction is one of the smallest register subcorpora of the TEC, meaning that the sample may not be representative of this register within Textbook English as a whole. Of course, this limitation is true of all the observations made on the basis of the TEC but, of the three textbook registers examined in 6.3.2 and 7.3.2, sparse data is most likely to affect the results concerning Textbook Fiction. Of the 42 textbook volumes of the TEC, some include no or very few fiction texts. In particular, the three French textbook series hardly feature any fiction texts (see Le Foll 2018a) which is worth highlighting because, when differences across textbook series *were* observed in any of the four analysis chapters, these almost always concerned the French textbook series (even though these differences rarely reached statistical significance at a threshold of $p < 0.01$).

This brings us to another aspect of the second set of research questions formulated at the beginning of Chapter 6 and restated in 7.1: *To what extent are (some of) the observed patterns of variation within Textbook English moderated by textbook series or their country of use?* This question goes back to one of the overarching research questions, which asked whether secondary school pupils in France, Germany and Spain are exposed to qualitatively different language input via their English textbooks (see 3.2.1). As has just been mentioned, hardly any significant differences were observed between the language of the nine different textbook series of the TEC – though some minor differences were occasionally observed between the nature of the language presented in textbooks used in France compared to those used in Germany and Spain. One major difference, however, has not yet been taken into consideration. Though it concerns the *quantity* rather than the *quality* of students' textbook-based language input and is therefore not directly relevant to the present study, it is worth remembering that the French subcorpus of the TEC is the smallest of the three "national" subcorpora (see Fig. 1 in 3.3.1.4). Moreover, as shown in Fig. 84, a large proportion of the language content of these textbooks is actually in French (labelled here as 'foreign', see 3.3.1.4). Another large chunk is made up of individual words and

sentences (mostly in the form of exercises). In other words, the French series of the TEC feature considerably fewer cohesive texts and, in particular, much shorter ones than in their counterparts used in German and Spanish secondary schools (see Fig. 2 in 3.3.1.4 for comparison).

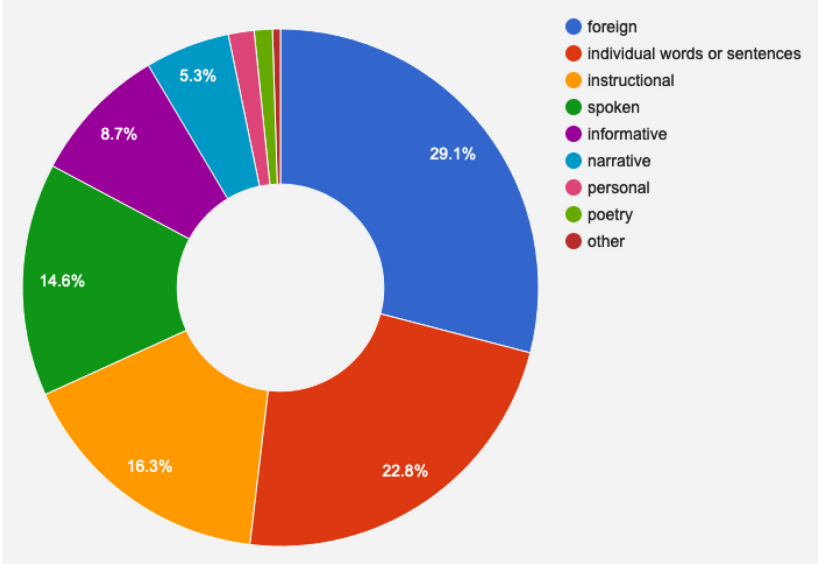


Fig. 84: Proportion of tokens in each register category of the French subcorpus of the TEC (as displayed by Sketch Engine; Kilgarriff et al. 2014)

Hence, although the TEC was constructed as a balanced corpus of textbooks with three series from each country (see 3.3.1.4), the number of texts and words are not distributed equally across each series or country of use. However, wordcount totals are difficult to compare across entire textbook series. This is because, on the one hand, one of the textbook series used in Spain (*Solutions*) and one of the French ones (*Piece of Cake*) only feature four rather than five volumes and, on the other, because the TEC does not include all the transcripts of additional audio/video materials for one German and one Spanish series (the older version of *Green Line* and *Achievers*) (see Table 5 and doi.org/10.5281/zenodo.4922819 for details). Hence, in Fig. 85, the distribution of words in the TEC is visualised per textbook volume. The plot shows the number of words (as calculated by the MFTE) in the Conversation, Fiction, Informative, Instructional, Personal correspondence and Poetry texts of the TEC (i.e., those used in the intra-textbook MDAs in 6.3.1 and 7.3.2). The bars are sorted by textbook proficiency level since we can expect the number and length of texts to increase as learners become more proficient in English. This is, indeed, what we observe in Fig. 85. However, the plot also makes evident that there are major differences in terms of the quantity of English that students are (potentially) exposed to via their textbooks.

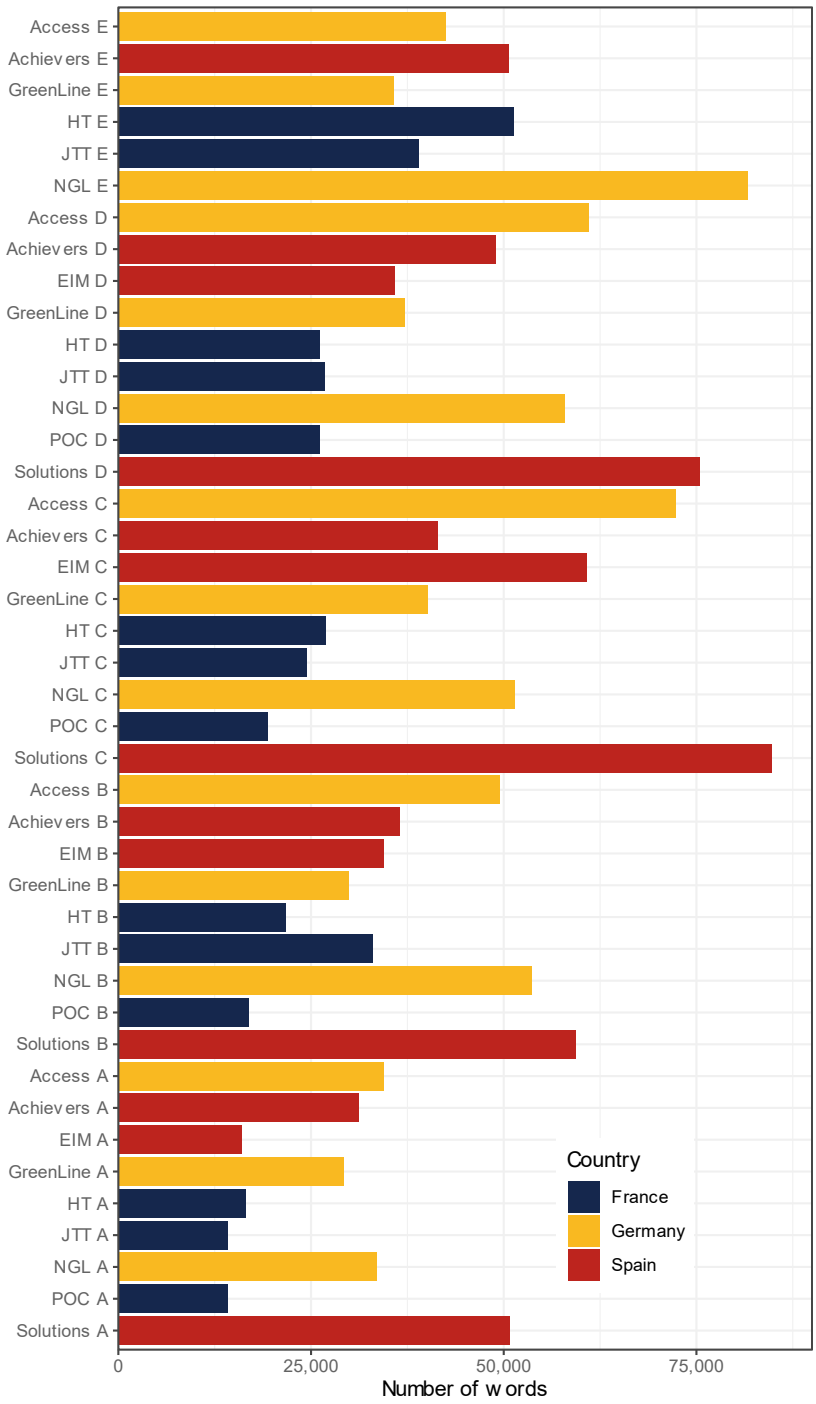


Fig. 85: Cumulative word counts of the Conversation, Fiction, Instructional, Informative, Personal correspondence and Poetry texts for each of the 42 textbook volumes of the TEC

Of course, it goes without saying that not all secondary school pupils learning with these textbooks will engage with every single text featured in the coursebooks; it is perfectly conceivable that some teachers use textbooks more as "sourcebooks" from which to pick and choose the texts and activities they believe to be most suitable for their students (see, e.g., Möller 2016; Schaer 2007). Equally, we can expect many teachers to supplement textbook materials with additional (textbook or non-textbook) materials. Nonetheless, it is quite striking that, across the textbook

proficiency levels A to D (corresponding to *collège* in the French educational system, see Table 3), the three French textbook series contain, in terms of *quantity*, considerably less English-language input than the series used in the German and Spanish educational systems. However, as stated above, all of the results presented so far indicate that the *quality* of this language input is very similar across the three countries represented in the TEC. It is not within the scope of the present thesis to investigate the pedagogical impact of either the quantity and quality of textbook-based language input; however, it is clear that both of these factors, together with a whole host of equally relevant teacher-, learner- and environment-based factors, will undoubtedly mediate learning outcomes.

The final research question formulated at the start of this chapter (7.1) was concerned with identifying the linguistic features that characterise Textbook English registers as compared to situationally similar, naturally occurring texts. Section 7.3.2.1 showed that Textbook Conversation is characterised by a highly nominal style. Its texts are defined by high lexical density and diversity. Hence it displays more diverse and often complex vocabulary than everyday real-life conversation. At the same, however, the results of the second PCA-based model showed that it features far fewer syntactic structures that are characteristic of more complex utterances, e.g., *that* and WH-subordinate clauses and causative subordination (see 7.3.2). Confirming previous analyses (see 5.3.5), phrasal verbs also appear to be conspicuously rare in textbook dialogues, including those featured in the more advanced textbook volumes of the TEC.

Section 7.3.2.3 demonstrated that, in the more advanced textbooks of TEC, informative texts share more similarities with the texts of the Info Teens corpus than differences. Nonetheless, Textbook Informative is, on average, characterised, on the one hand, by shorter words and fewer verbs belonging to the semantic category of existential/relationship verbs (e.g., *INVOLVE, IMPLY, REPRESENT, SEEM*) and, on the other, by higher relative frequencies of first-person singular and second-person references, contracted and negated verbs. Together with the results of the previous chapter, this suggests that some of the informative texts of the TEC can be considered to represent hybrid registers: their linguistic characteristics frequently cross the line between factual and impersonal informative writing, narrative explanations and involved, interactional communication.

As recommended by Lee (2000: 393; see also Biber 1990), the two PCAs that form the basis of the multi-dimensional models presented in this chapter were replicated on random subsets of the data to test the stability and robustness of the presented results. Aside from the inversion of the negative and positive ends of some of the components (which are entirely arbitrary), the results proved to be very stable across the six attempted replications (with randomised subsets of two-thirds of the data) of

each model (the code to run additional subset replications is included in the Online Appendix 7.7). The results of the first replication run on such a random subset of two-thirds of the data are visualised in Fig. 86 and 87 (printed in the following chapter). The plots can be compared to Fig. 74 and 76 (from the present chapter) that were computed on the basis of the full dataset. The same trends observed and described in 7.3.2 are evident. In addition to these randomised subset analyses, for each of the two models presented in this chapter, three additional PCAs were conducted using only the texts stemming from 1) the French, 2) the German and 3) the Spanish subcorpora of the TEC (see 3.3.1.1). Only minor differences in the ranking of the feature loadings were observed. The general trends can therefore be said to be stable across all three "national" subcorpora of the TEC which, again, confirms the finding that the language of lower secondary school EFL textbooks is very comparable across the nine series used in France, Germany and Spain that constitute the Textbook English Corpus. Together, these multiple replications testify to the robustness of the methodology described in 7.2.

Given that the results of the present chapter largely corroborate those of the previous chapter, it may be tempting to conclude that some of the additional steps described in this chapter's methodology (7.2) are not entirely necessary. This would, however, be a hasty conclusion on several grounds. For a start, one of the major advantages of the proposed revised framework lies in the replicability and robustness of the results it yields, which can be traced back to the choice of meaningful normalisation baselines for each linguistic feature (7.2.4), the (partial) deskewing of the distributions of these relative frequencies, the exclusion of features absent from a considerable proportion of texts or with very low communalities, and the removal of outlier texts (7.2.6). That said, as argued in 6.5, for a quick, preliminary analysis, conducting an additive MDA using the free and user-friendly MAT tool (Nini 2014) may be sufficient (see also Le Foll 2021b: 29–33). If, however, Biber's (1988) dimensions and the data underlying it are not considered (entirely) appropriate benchmarks to compare the language of 21st century secondary school EFL textbooks to (see also 6.5), then the present methodological framework is, in many ways, both considerably more robust and easier to implement than Biber's original method. For instance, using PCA rather than EFA as the underlying dimensionality-reduction method means that the largely arbitrary decision as to how many dimensions to extract can be linguistically and/or pragmatically motivated without fear of influencing the results of the entire model (see 7.2.5). In addition, not only does doing away with any thresholds to calculate dimension scores make theoretical sense (see 7.2.7), it also means that the output of the standard PCA functions offered in most statistical software can be interpreted without any additional manipulations. Finally, many statistical packages for PCA (including many open-source packages, see 7.2.10 and code in Online Appendix 7.2) also include multi-dimensional plotting functions, which means that visualising the results of PCA-based MDAs across several dimensions is comparatively easy. As we

have seen, visualisation is crucial to identifying internal patterns of variation (see 7.2.8 and Neumann & Evert 2021).

On the whole, the extensive use of visualisation has also increased the interpretability of the results. Some may argue that interpreting the feature loadings of the resulting models is more complex given that the frequencies entered in the PCAs represent signed logged-transformed standardised frequencies that were normalised to different baselines (see 7.2.4). However, the counts entered in the traditional MDA framework are not immediately obvious either: they represent standardised *z*-scores. Moreover, the linguistic motivations for opting for different normalisation baselines are considerably stronger than any interpretability arguments – especially given that the alternative (i.e., using word-based baselines) is certain to lead to genuine interpretation issues. In 7.2.4, we saw that, for example, when the relative frequencies of features such as contractions, negation and present tense per 1,000 words all contribute to one pole of a dimension, whilst the frequencies of nouns and prepositions per 1,000 words contribute to the opposite end, there is a risk that the dimension in question represents nothing more than the proportion of verbs to nouns. In other words, relying on frequencies normalised with a word-based baseline adds a considerable amount of uncontrolled variation to the relative frequencies entered in the analysis. This, in turn, can lead to difficulties in successfully conducting replications on new data that support the original conclusions of such studies (Wallis 2020: 74). By contrast, various tests have shown that the two models of Textbook English presented in this chapter are replicable on different subsets of the data which allows us to be relatively confident about their robustness. The potential pedagogical implications of these models, in combination with the cumulative understanding of Textbook English gained in the previous analysis chapters, are outlined in the following concluding chapter.

8 General discussion and implications

Language cannot be invented, it can only be captured.
❖ John McH. Sinclair (1997: 31)

This concluding chapter brings together the findings of the four analysis chapters (Ch. 4–7) and, in doing so, 8.1 attempts to answer the four sets of overarching research questions formulated in 3.2.1. In light of these findings, 8.2 considers the practical, pedagogical implications of the present study. First, 8.2.1 presents various ways to improve the representations of conversational English in ELT materials. Next, 8.2.2 outlines recommendations to improve the quality of the informative texts presented in these materials. This is followed by a section explaining how "bring[ing] textbooks for teaching English as a foreign language into closer correspondence with actual English" (Mindt 1996: 247) necessitates a register approach to ELT. This approach is explained in 8.2.3. Its practical consequences are broken down into implications for teacher training (8.2.4) and the design of EFL textbooks and other ELT materials (8.2.5). In spite of its pedagogically valuable results, the present study is clearly not without its limitations. These, together with methodological implications for future textbook language research, are discussed in 8.3. To conclude, future research avenues are outlined in 8.4 before summarising this study's main take-home messages in 8.5.

8.1 Summary and conclusions

The present study set out to describe the language that secondary school pupils in France, Germany and Spain are exposed to via their textbooks. As explained in Chapter 1, textbooks constitute one of the major, if not the most important, vector of English language input that EFL learners encounter in the first four to five years of their secondary schooling. Although it is popular knowledge that the language portrayed in EFL textbooks somehow "feels" different to the kinds of English used outside the classroom, this study is the first to attempt to model the nature of these linguistic peculiarities across different registers and proficiency levels by accounting for many different linguistic features and their co-occurrences.

The literature review carried out in Chapter 2 showed that, to date, the representations of a broad range of individual linguistic features have been examined in EFL/ESL textbooks. In 2.1 some of these studies were described as 'intra-textbook analyses' because they seek to explore and describe the language of EFL textbooks without relying on any comparison benchmarks. By contrast, comparative textbook

language analyses draw on reference corpora or corpus-based lists to infer what is different or special about the language of textbooks. In this context, two main issues were identified. First, explorations of how individual lexico-grammatical features are represented in EFL textbooks had, up until now, failed to account for potential interactions between these features and, second, Textbook English studies had largely ignored the fact that textbooks represent a range of different registers. Thus, much textbook language research had (often implicitly) assumed that Textbook English constitutes a homogenous variety of English with no (systematic) sources of internal variation. The present study has shown, across all four analysis chapters, that this assumption is not justified. By far the largest source of intra-textbook linguistic variation was observed across the different modes and registers represented in the textbooks of the TEC. In addition, textbook proficiency levels were also found to be associated with significant patterns of intra-textbook variation. By contrast, no major patterns of linguistic variation could be attributed to the idiosyncrasies of groups of textbook authors and/or editorial policies.

The first analysis chapter (Chapter 4) presented an extended conceptual replication of Römer (2005: Ch. 5 and 6), in which representations of the progressive in the spoken-like components of two series of EFL textbooks used at lower secondary level in Germany were compared to a corpus of spoken British English. Römer's (2005: Ch. 5–6) analyses were replicated using data from the Conversation subcorpus of the TEC, which includes more recent versions of the same textbook series that comprised Römer's textbook dialogue corpus (see 4.1.2). Many of the trends observed in the original study were found to be still significant in the dialogues of contemporary secondary school EFL textbooks. For instance, past-tense progressives, progressives with contracted auxiliary verb forms and negated progressives remain under-represented in many contemporary textbook dialogues. The replication carried out in Chapter 4 can be said to be 'extended' for three reasons. First, it examines the progressive in the language of EFL textbooks used in France and Spain, as well as Germany. Second, in addition to conversation, Chapter 4 also reports on morpho-syntactic analyses of the progressive in a second register of school EFL textbooks, namely: fiction. Third, the original study was expanded to include in-depth lexical and semantic analyses of the verb lemmas associated with the progressive in both Textbook Conversation and Textbook Fiction, thus adopting a truly lexico-grammatical approach to the study of the progressive (see 4.3.4). This was achieved using collostructional (see 4.2.3) and correspondence analyses (4.2.4). The additional comparison of representations of the progressive in Textbook Fiction with those found in the reference Youth Fiction corpus indicated that many of the reported linguistic peculiarities concerning the use of the progressive in Textbook English are actually only true of representations of spoken, conversational English in textbooks. However, some differences were observed in Fiction, too – for instance, concerning the functional use of the progressive in 'framing' contexts (see 4.3.3.5).

Lexico-grammatical/phraseological analyses were also at the heart of the second case-study chapter, Chapter 5, which attempted to "make sense of MAKE" in the same two registers of school EFL textbooks as in Chapter 4: Conversation and Fiction. In many ways, the use of MAKE in these two subcorpora of the TEC was found to be very comparable to how MAKE is used in the reference Spoken BNC2014 and Youth Fiction corpora. However, Chapter 5 also pointed to some disconcerting misrepresentations of the use of MAKE in the textbooks of the TEC. In particular, the representations of phrasal verbs (e.g., MAKE *out* in the sense of PERCEIVE, MAKE *sth. up* in the sense of INVENT) and collocations pertaining to the semantic field of discourse and communication (e.g., MAKE *a(n) argument/assumption/complaint/point*) were found to be particularly weak.

Having explored two sets of specific lexico-grammatical features in two contrasting textbook registers in these two case-study chapters, Chapter 6 paved the way for a multi-feature description of Textbook English. To this end, the texts of the Conversation, Fiction, Informative, Instructional, Personal correspondence and Poetry & rhyme subcorpora of the TEC (see 3.3.1.4) were first mapped onto the six dimensions of variation that constitute Biber's (1988) multi-feature/dimensional model of General Spoken and Written English. The results reported in Chapter 6 make clear that register differences must be accounted for when describing and/or evaluating the language of school EFL textbooks. On Biber's (1988) Dimension 1, which accounts for the most variance across the texts of the TEC, register was shown to explain by far the largest proportion of the variance observed in dimension scores (63%). By contrast, textbook series accounted for just 5% of the variance. On this same dimension, the different proficiency levels of the textbooks only explained 2% of variance; however, this percentage was found to be considerably higher on some of the other dimensions of Biber's (1988) model. Significant interactions between specific registers and target learner proficiency levels were also observed. These were found to be particularly relevant in the description of Textbook Fiction: on average, the narrative texts of the TEC are very different to those found in real novels; yet these differences are almost entirely driven by proficiency level effects caused by beginner textbooks relying on present-tense narration in the units preceding those introducing the past tense (see 6.3.2.2 and 7.3.2.2).

In spite of many interesting and pedagogically relevant observations, Chapter 6 concluded that not all of Biber's (1988) dimensions can contribute to meaningfully describing the language of secondary school EFL textbooks. Biber's (1988) Dimensions 4, 5 and 6, in particular, were shown to be of very limited use because they are made up of just a handful of features, which are rare in general English and even rarer in Textbook English. Further methodological issues were identified both in the selection of linguistic features that Biber's (1988) model relies on and their operationalisations. Some of the features and feature operationalisations were shown

to be unsuitable for the present data (e.g., nominalisations, stranded prepositions, discourse markers, etc.) (see 6.3.2.1). In some cases, this was due to operationalisations relying on punctuation marks – which, aside from question marks, are entirely absent from the transcripts of the Spoken BNC2014 corpus (see 3.3.2.1).[60] Other issues discussed in 6.5 concern the statistical methods and parameters that the traditional MDA framework relies on.

As a result of the findings of Chapter 6, suggested improvements to the MDA framework were presented in the first half of Chapter 7 (7.2) and implemented in the second half (7.3). The first PCA conducted as part of this improved MDA framework captures intra-textbook variation on four dimensions. It models Textbook English as a variety of English that, like all regional and domain-specific varieties of English, varies considerably across different registers. Two main dimensions of register variation were identified: the first ('Overt instructions and explanations', see 7.3.1.1) clearly delineates instructional texts and instructions from the rest of the registers featured in secondary school textbooks, whilst the second ('Involved vs. Informational Production', see 7.3.1.2) draws a continuum between involved, interactional speech and informational written language with textbook dialogues at the top of the scale and informative texts at the bottom. As such, this second dimension is very reminiscent of Biber's (1988) Dimension 1 (see 6.1.1). A third dimension of linguistic variation ('Present/factual vs. Past/speculative', see 7.3.1.3) explains some differences between the fictional texts of the TEC and the remaining text registers, whilst a fourth dimension ('Clausal complexity', see 7.3.1.4) highlights further discrepancies between Textbook Conversation and the Spoken BNC2014, though these two dimensions account for considerably less variance than the first two.

In 7.3.2, a second PCA was conducted – this time with the aim of modelling three major registers of Textbook English in relation to the three corresponding target learner language reference corpora: the Spoken BNC2014 (3.3.2.1), Info Teens (3.3.2.2) and Youth Fiction (3.3.2.3). Here, too, the results confirm the observations made in the two case-study chapters (Chapters 4 and 5) and the one applying additive MDA (Chapter 6): register differences within textbooks do exist but are considerably less pronounced than in 'real-life' English. Similarities and differences between these three textbook registers and comparable 'real-life' registers were modelled on four dimensions of linguistic variation.

Both Chapters 6 and 7 have demonstrated that MDA is highly suitable for describing Textbook English and, in combination with mixed effects linear models, an effective way of disentangling the various sources of variation and their interactions within

[60] Note that although it is true that punctuation marks were added at utterance boundaries for the novel MDAs carried out with the MFTE in Chapter 7, they only served to boost the accuracy of the underlying part-of-speech tagging process (see 3.3.2.1).

Textbook English. In 7.2.10, it was noted that some have raised concerns that, as a methodology, MDA has replicability issues. Given that the revised MDA framework presented and applied in Chapter 7 is relatively novel, it was all the more important to ensure that the results it yielded were robust. This was achieved by successfully running replications of both models presented in Chapter 7 on various subsets of the data (see 7.4).

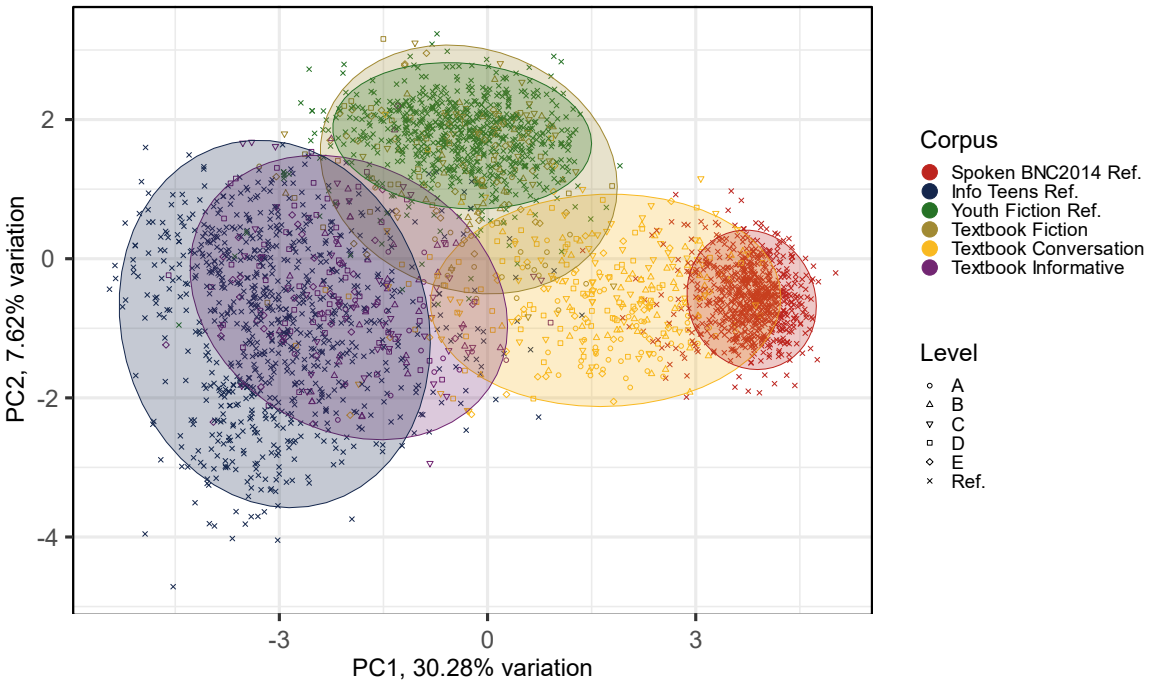


Fig. 86: Projection of texts on PC1 and PC2 from a random 2/3 split-data analysis of the three subcorpora of the TEC and the three reference corpora

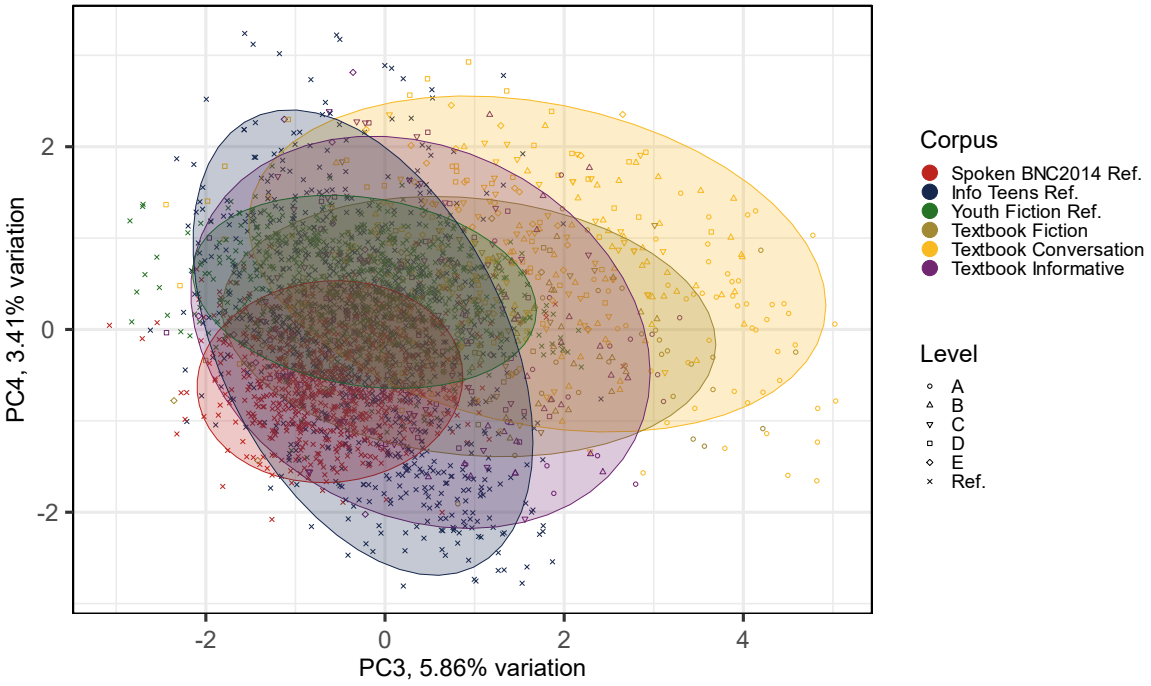


Fig. 87: Projection of texts on PC3 and PC4 from a random 2/3 split-data analysis of the three subcorpora of the TEC and the three reference corpora

Fig. 86 and 87 illustrate the robustness of the results. Though they may, at first sight, appear identical to Fig. 74 and 76, they, in fact, visualise the results of a model computed on the basis of a random subset of two-thirds of the data (see Online Appendix 7.7). The trends discussed in 7.3.2 are clearly identifiable on both these figures. In light of the success of these replicability and robustness tests, the following section provides answers to the four overarching research questions formulated in 3.2.1.

1. How homogenous is Textbook English as a variety of English? Which factors mediate intra-textbook linguistic variation?

The results of the four analysis chapters converge to confirm that the language of school EFL textbooks cannot be conceptualised as a homogenous variety of English. The results of the PCA-based model of intra-textbook variation in 7.3.1, in particular, showed that both register and textbook proficiency level effects are drivers of intra-textbook linguistic variation; yet register accounts for considerably more of this variation. Furthermore, many significant interactions between different registers and textbook proficiency levels were observed and quantified.

2. To what extent are French, German and Spanish secondary school pupils confronted with varying English input via their textbooks?

Given that the textbooks examined as part of this project are designed for and/or used in very different educational contexts, it was hypothesised that there would be noticeable differences in the linguistic nature of the language input that students learning English in France, Spain and Germany are exposed to via their textbooks. However, throughout the four analysis chapters, no systematic differences could be discerned between the language of textbooks used in France, Germany and Spain (though the French textbook series were found to deliver substantially less language input, see 7.4). Indeed, even when comparing the nine textbook series of the TEC, very little of the observed intra-textbook linguistic variation could be attributed to the authors or editorial policies of individual textbook series. Thus, in answer to this second overarching research question, we can conclude that the nature of the language that French, German and Spanish secondary school pupils are confronted with via their textbooks is, in fact, remarkably similar, in spite of different academic traditions, school systems and textbook publishing traditions. Possible reasons for this will be discussed in the following sections.

3. To what extent is the language of current EFL textbooks used in secondary schools in France, Germany and Spain representative of 'real-life' English as used by native/proficient English speakers in similar communicative situations? To what extent are some registers more faithfully represented than others?

Except for the rare few narrative texts in beginner textbooks which rely on present-tense narration and understandably feature very limited vocabulary, all of the results converge to show that fiction is the most faithfully represented register in school EFL textbooks. This is demonstrated by the large overlaps between the Textbook Fiction and the Youth Fiction ellipses on most projections of texts (see, e.g., Fig. 86 and 87 in this chapter). Large overlaps can also be seen in the Informative register. By contrast, on these multi-dimensional projections of texts, only a small proportion of the Textbook Conversation texts are situated within the ellipses of the Spoken BNC2014 – thus pointing to major differences between natural speech and textbook representations thereof. The results of the two case-study chapters also confirm these trends. Section 8.2.1 is devoted to the potential implications of inauthentic representations of conversational English in pedagogical materials.

4a. Which (clusters of) lexico-grammatical features are characteristic of Textbook English?

The clusters of lexico-grammatical features that characterise the various registers of Textbook English were discussed at length in the preceding four analysis chapters. To summarise, across different registers, the linguistic peculiarities of Textbook English are perhaps best illustrated in the projection of texts on the first and second dimensions of the second PCA-based model conducted in 7.3.2 (see Fig. 74). Whilst it shows that many Conversation, Fiction and Informative texts featured in school EFL textbooks are, at least on these two dimensions of linguistic variation, very akin to or even practically undistinguishable from situationally similar real-life texts, the ellipses of the TEC registers are nonetheless noticeably shifted towards the middle of the plot as compared to those of the corresponding reference corpora (see Fig. 74). A further gap between textbook and non-textbook language was observed on the third, 'Pedagogically adapted vs. Natural', dimension that emerged from the second PCA-based model of Textbook English: the texts of the TEC tend to be shifted towards the positive end of this dimension (see Fig. 87). In other words, the most prototypical exemplar texts of Textbook English are likely to be located around zero on the 'Spontaneous interactional vs. Edited informational' and 'Narrative vs. Non-narrative' dimensions and towards the 'adapted' end of the 'Pedagogically adapted vs. Natural' dimension. Excerpts (244)–(246) are examples of such texts and can thus be said to be "prototypically textbook-like". Although (244) corresponds to an informative text, (245) represents fiction, and (246) a conversational interview, taken

from three different textbook series, they are stylistically remarkably similar. They share many linguistic features: an abundance of nominal phrases, high lexical density, high relative frequencies of *yes/no* and WH-questions, a strong preference for the modal *can* over other modal verbs and BE as a main verb over other lexical verbs, to mention but a few. All in all, the texts are written in an informal style that, in many ways, appears to emulate some aspects of casual conversation yet, as seen in all the comparisons with the Spoken BNC2014 in Chapters 4–6, lacks many of the defining features of natural, spontaneous conversation.

Excerpt (244) is a representative example of a typical textbook dialogue in that, whilst it attempts to model some natural disfluencies in the form of hesitations and the occasional paralinguistic filler or discourse marker (*mmm*, *well, yes*), it remains first and foremost a vehicle for vocabulary learning. Hence, the dialogue is built around words, phrases and idioms that are to be acquired by the learners. In excerpt (244), these include *playground, kangaroo, palm tree, suit sb. down to the ground, kookaburra, sailing boat, bush, rollers,* etc. In such texts, the textbook authors clearly prioritise the placement of these vocabulary items over the naturalness of the dialogue. This is not to say that there is no valid pedagogical justification to do so. Potential issues only arise if learners are expected to acquire the necessary linguistic and pragmatic competences to interact with others in spontaneous conversation solely on the basis of such, highly unnatural, models of spoken interaction (more on this in 8.2.1).

(244) Now, **what** about you, **Charlene**? **What** it is like to live here**?**
Well, it's really great! We have a vast **garden**. And it'**s** a great **playground** for **Joey**, our two-year-old **kangaroo**. The **garden**'**s** full of **palm trees** and exotic **plants**. There'**s kookaburras** and other **birds** that sing all **day** long. It'**s** really wonderful! Our **house is** typically Australian, which suits me down to the **ground**. And the **view** from the front **porch is** fabulous! We **can** see **sailing boats** everywhere in the **summer**.
Now, tell us about your **everyday life**. **What** do you do here in **Sydney?**
On **Sundays**, we usually go walking in the **bush** with the whole **family**. And we also go shopping at the **Rocks** or on **Circular Quay**. And if we've got **time**, we go for a **walk** in **Hyde Park**.
Charlene?
We often go on **visits** to the **Aquarium**, **Darling Harbour** and **Taronga Zoo** with **friends**. We go to **Bondi beach**. Mmm... there'**s** great **rollers** there for **surfing**. I love going to the **beach** after **school** when it'**s** hot, in **January**... Yes... And all the **family** goes jogging every **morning**, even me... by the **sea** and, well, it'**s** just fantastic to be out, out in the good **weather**. It'**s** not so cold around here. <TEC: New Bridges 2[de]>

By contrast, and although it is cast as an informative text in the form of a short article, text (245) is much more spoken-like than the majority of informative texts found in the Info Teens corpus: it features rhetorical questions, many second-person references and a number of discourse markers typical of speech (e.g., *well*).

(245) The British **soap opera Coronation Street** started in 1960 and **is** still on TV today. Other popular **soaps are Neighbours**, **Emmerdale** and **EastEnders**. All the **soaps** try to be realistic about **life** with its happy **times**, its **problems** and some **violence**. The **name** "**soap opera**", or just "**soap**", goes back to **radio dramas** in the 1930s - the **commercials were** for **housewives**, and they advertised **soap**, and other cleaning **products**. Want to be a **star?** Want to be discovered**?** Not so fast! Before you **can** get anywhere, the **programme** has to "cast" you first. Have you ever been invited to do a **casting?** You haven't**?** Well, **TEENBUZZ** tells you all about it. **Matt Stirling** from **EastEnders can** give you a few **tips**, too. First, you talk to an **agent** and give him or her your **photo**. Then one day the **agent** is phoned by a **Casting Director** who is looking for a special **character** for a **soap**. She tells the **agent** who she needs. Let's call him "**justin**". So the **agent** looks through his **files** and finds a **photo** of - you! Your **photo** is sent to the **Casting Director**, who looks at hundreds of **photos** for the right "**justin**". She likes your **face**! <TEC: Green Line 3>

Excerpt (246) is a representative example of a fictional text from a textbook targeted at learners in their second year of English classes at secondary school. It is narrated in the present tense and features more contracted verbs and occurrences of BE as a main verb per finite verb phrase than the average novel targeted at teenagers and young adults. It would be inappropriate to conclude that it is "unnatural" or "inauthentic" simply by virtue of being situated among the clusters of "prototypically textbook-like" texts on Fig. 74 and Fig. 87. The fact that such texts are amongst the most stereotypically "textbook-like" texts located in the middle of Fig. 86 and 87 and towards the positive end of Dimension 3 (see Fig. 76) merely points to the narrative nature of prototypical textbook texts, regardless of the register they intend to portray, as demonstrated in (244) and (245).

(246) "There **are** good ideas and bad **ideas**," thinks **Ruby**. "And this **is** a bad **idea**." On **Saturdays**, **Ruby** usually reads or paints at **home**. She often goes out and takes unusual **photos** of **things** in her **town**. Sometimes she even makes short **films** and uploads them onto her **website**. But today, on this cold, sunny **Saturday** in **April**, **Ruby** is running up a **mountain**. OK, it**'s** a very small **mountain**. But **Ruby doesn't** like **mountains**. And she hates **running**. **Ruby** and her **friends** are raising **money** for a **charity**. They want to help **schools** in **Africa** buy new **computers**. And yes, **Ruby** knows it**'s** a good **idea** to raise **money** for **charity**. But **running**? Up a **mountain**? That **is** simply terrible. **Ruby** stops, closes her **eyes** and holds her **head** in her **hands**. She feels terrible. She **is** out of **breath**, her **chest** is burning and her **legs** hurt. "Next **time**, ... next **time** I **can** sell a **picture**... or clean **cars**... give away all my **money**... anything! But I**'m** never, never doing this again!" She sits down on a **rock**. The air **is** cold and her **breath** forms small white **clouds**. The **sun** is shining brightly in a clear blue **sky**. <TEC: Achievers A2>

4b. To what extent are these defining features stable across entire textbook series? To what extent are some specific to certain proficiency levels?

The present study has made clear that many, but not by any means all, linguistic features that are highly typical of Textbook English are stable across entire textbook series – in other words, across the different proficiency levels represented in the TEC. However, proficiency level effects were also found to interact with how linguistically similar the Conversation, Fiction and Informative texts featured in secondary school EFL textbooks are to situationally similar texts that students are likely to engage with outside the classroom.

Given the obvious relationship between the language of foreign language textbooks and the targeted proficiency level of the learners, it is not surprising that we find such proficiency level effects on many of the extracted dimensions of variation in Textbook English. The strongest of these interactions were observed in the Fiction register, followed by Informative texts and, lastly, Conversation. Curiously, however, the proficiency level trends observed in the Conversation subcorpus of the TEC follow, on almost all the dimensions of variation scrutinised, the opposite direction to what common sense would have predicted: Conversation texts from the most advanced textbooks of the TEC are, on average, the ones that are most different to the transcripts of the Spoken BNC2014. Some of the potential pedagogical implications of this and the other key findings summarised above are discussed in the following section.

8.2 Pedagogical implications and recommendations

Before beginning to reflect on the potential pedagogical implications of the present study, it is important to remember that all of the analyses presented in this thesis are, by nature, descriptive and exploratory. The present study did not attempt to investigate the *effectiveness* of Textbook English, in other words the extent to which prototypically textbook-like language may or may not support EFL learners in their English acquisition processes. It merely attempted to provide a comprehensive description of the language of lower secondary school textbooks used in France, Germany and Spain and, in doing so, to raise awareness of what constitutes Textbook English.

In terms of pedagogical implications, it should be stressed that the proficiency level effects that were found to mediate linguistic variation within Textbook English suggest that some of the identified specific characteristics of Textbook English are pedagogically well-founded, or, at the very least, intended by the textbook authors and editors. Indeed, it certainly would not make pedagogical sense to expose learners to the same kind of language in their first year of learning English as in their fifth.

At the same time, this study has demonstrated that, in many ways, Textbook English does represent a distinct variety of English.

When Segermann (2000: 339) asserted that foreign language textbook texts represent a text type "*sui generis*", she probably had the kind of texts situated in the middle of Fig. 74 in mind, e.g., (244)–(246). Indeed, excerpts (244)–(246) epitomise the type of contrived texts which have been meticulously crafted by textbook authors in order to fulfil very specific pedagogical criteria. Constraints on textbook authors are multiple. In many cases, these texts are largely determined by the lexis and grammar of the stringent learning progressions that are enshrined in the textbooks' tables of contents and which are, in turn, devised on the basis of the curriculum and syllabus in force (see, among others, Burton 2019; Gießing 2004: 84; Quetz 1999: 16–17; Segermann 2000: 339). It is not uncommon for textbook authors and teachers to believe that this particular type of contrived text is – at least in the first few years of language learning – indispensable because learners are only to be exposed to "controllable portions" of language (Segermann 2000; see also Thornbury's [2000] infamous "grammar McNuggets").

This belief can be traced back to Pienemann's (1984) teachability hypothesis that postulates that, when acquiring new morpho-syntactic structures, L2 learners follow a natural developmental sequence which means that they can only acquire a new structure once they have mastered the structures that precede it in this developmental sequence (see also Krashen 1982; 1985 on the related notion of "comprehensible input"). The hypothesis of universal developmental sequences has been, to some extent, confirmed in empirical studies (e.g., Ellis 1989; Spada & Lightbown 1999). What has been refuted, however, is the accompanying hypothesis that teaching a structure for which learners are not yet "developmentally ready" will not only fail to result in learning, but also have a detrimental effect on learning outcomes (e.g., Larsen-Freeman 2006; 2018; Spada & Lightbown 1999; Ur 2011: 513). Yet, this underlying belief that learning a foreign language is a linear process and that a text is only accessible to learners if they are already familiar with every aspect of its lexis and grammar remains widespread (see, e.g., Phakiti & Plonsky 2018).

Given that these pedagogically contrived texts (and, again, it is worth reiterating that these do not, by any means, represent *all* textbook texts!) are created, first and foremost, to meet very restrictive criteria, they cannot be expected to "sound natural" or to be perceived by learners as such:

> Von ihr [dieser Textsorte *sui generis*] zu verlangen, dass sie von einem Schüler wie ein normaler Text rezipiert wird (mit normaler Erwartungshaltung, Eigeninteresse und entsprechender Bereitschaft zur selbsttätigen Sinngebung), hieße, ihre Funktion zu verkennen und die Quadratur des Kreises zu verlangen. [To expect that it [this kind of textbook text *sui generis*] be received by pupils as a normal text (with normal expectations,

intrinsic interest and hence the willingness to make sense of it by themselves) would be to misjudge its function and to demand the squaring of the circle.] (Segermann 2000: n.p.).

Indeed, if textbook dialogues are often the butt of *Brian is in the kitchen*-type jokes (see Chapter 1), it is precisely because learners (and their parents) are well aware that many of these dialogues are anything but natural-sounding, as demonstrated in the following extract from a focus group interview conducted with secondary school pupils in Germany:

> M: (...) Wenn irgendwie so ne englische Sendung ist im Fernsehen, da versteht man nicht jedes Wort, weil die umgangssprachlich reden und wir lernen in der Schule ja ein anderes Englisch. Wir lernen eher so das Englisch, äh, so
> M: normales Schulenglisch
> M: normal
> M: Schulenglisch
> M: Schulenglisch, und das ist anderes Englisch was die reden. Das ist umgangssprachlich, die haben andere Wörter
> M: da ist das auch viel schneller, unverständlicher
>
> [M: (...) When, like, in an English programme on TV, you don't understand every word, because they use colloquial language and in school we learn a different English. We kind of learn the English, ahm, like
> M: normal school English
> M: normal
> M: school English
> M: school English, and this is different from the one they use. It's colloquial, they use different words
> M: it's a lot faster, and not so easy to understand] (Grau 2009: 170, 173; translation Grau)

In German, the widely-used term *Schulenglisch* [school English] is mostly used in a pejorative sense to describe "a form of English that marks its users as having acquired the language in school" (Grau 2009: 170). It is considered different from the English that is actually used by competent English speakers. In the following extract of an interview, a Spanish L1 English L2 speaker refers to the same phenomenon as *inglés de libro* [(text)book English]:

> Yo cuando llegué aquí [en Inglaterra] por primera vez es que hablaba un inglés de libro y me sentía fatal. O sea, no tenía esos recursos de conversación más informal, más de registro. Pues sí, con amigos, con familia. Me faltaba ese vocabulario y digo llevo años estudiando inglés y era muy artificial. [As for me, when I arrived here [in England] for the first time, I spoke textbook English and I felt awful. Like, I didn't have those resources for more informal conversation, [that are more appropriate] for that register. Well, that is, with friends, with family. I lacked that vocabulary and, I mean, I'd been studying English for years and it was so fake.] (Pérez-Paredes & Abad forthcoming)

Here, too, textbook-based EFL instruction is seen as inadequate in terms of teaching spoken communicative skills. Indeed, the layperson's perception that representations of spoken English in EFL textbooks are particularly inauthentic has been confirmed in the present study. In both case-study chapters and on all dimensions of linguistic variation examined in the two MDA chapters, Textbook Conversation was

consistently found to be the least natural-sounding of textbook registers. This is why, in turning to the practical implications of the present study, the following section begins with the potential pedagogical implications of these unnatural representations of spoken English in secondary school EFL textbooks. Having pointed to what makes these dialogues potentially problematic, it goes on to suggest solutions to improve the quality of EFL materials intended to model conversational English.

8.2.1 Improving representations of conversational English

Chapters 6 and 7 showed that by far the greatest gap between Textbook English and natural English that learners are expected to encounter outside the EFL classroom is found in the Conversation register. To begin, however, it is worth highlighting that not all textbook dialogues were found to be strikingly different from naturally occurring conversation. Excerpt (247), for instance, is situated within the ellipses of the Spoken BNC2014 on all projections of texts visualised in Chapter 7. This is because, among other factors, it features relatively high frequencies of fillers and interjections, discourse markers, the modal *could*, contracted verbs, negation, and an abundance of first and second-person references.

(247) Amy: **Hi**, Nick.
Nick: **Hi**, Amy. Amy, **is** this **your** backpack on the floor?
Amy: That**'s** right.
Nick: **Well**, **could you perhaps** put **it** somewhere else? **It's kind of** in the way.
Amy: **No, it's not**. **It's** where **I** always **leave it**.
Nick: **Yes**, **I know you** always **leave it** there. **And it's** always in the way. This **is** a pretty small place, Amy. **So perhaps just** for once **you could** put **your** backpack somewhere where **it isn't** in the way, **hmm**?
Amy: **You don't own** this place, Nick. **So don't try** and **tell me** what to do. **I** came in early to get some things done. **I** put **my** backpack on the floor. **You deal** with **it**! <TEC: English in Mind 4>

Thus, it would be a grossly misleading simplification but to claim that all textbook dialogues are poor representations of spontaneous spoken language. That said, the majority of the texts of the Textbook Conversation subcorpus did score much lower than (247) and the texts of the Spoken BNC2014 on both Biber's (1988) 'Involved vs. Informational Production' dimension (see 6.3.2) and the functionally very similar 'Spontaneous interactional vs. Edited informational' dimension that emerged from the second PCA-based model in 7.3.2. Most worryingly from a pedagogical point of view, the statistical analyses of the models in Chapters 6 and 7 showed that the dialogues of the most advanced textbooks of the TEC are, on average, the least representative of natural conversation (see 6.4 and 7.3.2.1).

Depending on the learning objectives associated with these texts, sound pedagogical reasons may well justify the unnaturalness of some of these artificial dialogues. If a textbook dialogue's primary aim is to teach a pre-defined list of nouns, it may make

sense to construct the text around these nouns. If, however, the primary purpose of these texts is to develop learners' oral competences (i.e., their receptive and/or productive skills), then such low scores on oral–literate dimensions of variation are a cause for concern. Excerpt (248), for example, is an extract of a textbook dialogue for the fifth year of secondary EFL instruction that is situated closer to the edited/informational end of the 'Spontaneous interactional vs. Edited informational' dimension than any of the conversation transcripts of the Spoken BNC2014. In fact, it is linguistically far removed from natural conversation on all dimensions of variation explored in the present thesis. This is the result of its considerably higher type/token ratio and longer average word length than most real-life conversations, as well as the fact that it features many complex nominal phrases, which, in turn, lead to high relative frequencies of prepositions and attributive adjectives.

(248) Journalist: This is **Sally Gordon** here in **Leicester Square, London**. I'm right **in** the **middle** of **sports fans**. Excuse me, Sir. Who is your **favourite sports hero**?
Dwayne: Definitely, **Chris Hoy**, the **British track cyclist** - won two **gold medals**. He represents **strength** and **courage**, he never gave up.
Journalist: What about you? Who is the **best representative of** your **country**?
Donna: **Kobe Bryant** for sure. I'm American and we are very patriotic when it comes **to sport**. He has shown the **world** we remain the **dominant leaders in basketball**, no doubt. And **Michael Phelps** of course.
Journalist: Why?
Donna: Why? He has just won four **golds** and two **silver medals** and he is a **record holder**. The **dream** came true. Incredible. That's why he is nicknamed "the **Baltimore Bullet". He symbolises determination, generosity, hope... great values**. You see, he's a **role mode**l! He will be remembered forever. <TEC: New Mission 2de>

Hence, whilst textbook dialogues such as (247) expose learners to interactional, genuinely conversation-like language that they are likely to encounter outside the classroom, texts such as (248) cannot be considered to be realistic models for EFL learners to acquire spontaneous spoken language comprehension and/or production skills. As mentioned above, such texts can be argued to serve other pedagogical purposes, e.g., the high lexical diversity of excerpt (248) may be specifically aimed at increasing learners' receptive vocabulary range by introducing learners to many nouns from a single semantic field (e.g., *strength, courage, determination, generosity, hope*, etc.).

Given that such dialogues represent the norm rather than the exception in secondary school EFL textbooks, textbook authors and teachers ought to carefully consider the primary pedagogical purpose of such highly unnatural-sounding dialogues. If they are destined to model and enhance learners' conversational skills, they should consider replacing them with authentic materials. If no suitable authentic materials can be sourced, corpus data or the results of corpus-based studies may be used to revise

them. The following paragraphs illustrate these various options on the basis of excerpt (248).

Given the now widespread availability of corpora and corpus tools, there is no longer a need for textbook authors to systematically craft their textbook dialogues from scratch (for a list of freely available English corpora and corpus tools, see Le Foll 2021c: Appendix). Of course, this is not to suggest that the transcripts of spoken corpora such as the Spoken BNC2014 should be printed *verbatim*, as in (249), in coursebooks. However, if learners are to be expected to develop real-life conversational skills, the kind of natural interaction that is captured in such transcripts ought to be featured in at least some of the audio-visual materials that accompany EFL coursebooks. For instance, in a conversation about meeting a sports hero from the Spoken BNC2014, excerpt (249) demonstrates how speakers frequently interrupt each other, ask for clarification when things are unclear (e.g., *who is who is it?*) and manifest their interest in the conversation with laughter, other paralinguistic sounds and various interjections (e.g., *yeah*, *uhu*, *mm*, *oh cool*).

(249)
```
I met my hero didn't I? I text you remember?
yeah
who did you meet?
Hans no way Rey
you were[?]
who is who is it?
he's talking about a famous mountain climber
famous bike rider
oh okay
he's astonishing I went well I wasn't very old I think maybe I was
employed to drive maybe I was just starting to drive
uhu
and there was like a big there used to be a big like erm biking event in
<anon type="place"/>
mm
and he was in like a a show and the things he did honestly cos bikes
were crap in those days but he was
<unclear/>
amazing and he had the stereo going in the background and he like had
these picnic tables and he was like hopping onto them and jumping
between them
oh cool
yeah he had all quite
<vocal desc="laugh"/>
people would lie on the floor and he was like jumping over them and he'd
get his front wheel and then he'd put the front wheel onto there and
<unclear/> it was <trunc>am</trunc> a show like <BNC214: SBKN>
```

As demonstrated by the complex turn structure in (249), faithfully scripting such dialogues is no mean feat. For "cleaner", less "chaotic", yet linguistically accurate representations of spoken language (as empirically demonstrated by, e.g., Quaglio 2009; see also Werner 2021 on the relationships between the language of scripted

telecinematic language and face-to-face conversation), textbook authors and teachers may want to turn to film and TV series for more readily usable materials. Indeed, screenwriters are professionals in imitating natural speech. Often such materials can relatively easily be adapted for classroom use. Excerpt (250), for instance, was sourced from the freely accessible TV Movie Corpus hosted on the online corpus platform english-corpora.org (Davies 2019) by searching for the target collocation *gold* MEDAL[61] featured in the original textbook dialogue (248). It could be used with beginner learners of English without any modification.

(250) Hey, Smoochie, come here.
I have something pretty special to show you. Come here, bub, do you want to see a video of one of the best gymnasts in the entire world? It's not just me saying that. She won an actual **gold medal**. Her name is Simone Biles, that's her right there.
Oh, Simone Biles, she is amazing.
Right? Here, watch this. [Crowd_Noises_On_Laptop]
Wow, she is very flexible.
[…]
There we go, look at this, look at this.
Whoa!
Whoo! Whoo, she's so good, did you like that?
It was really good.
I know, right? <TV Movie Corpus: I'm Sorry (2017)>

In staying with the topic covered in textbook excerpt (248) representing unnatural conversational English, a quick query for the phrase *sports* HERO in the TV/MOVIES subcorpus of the *Corpus of Contemporary American English* (COCA; also available on english-corpora.org; Davies 2010) returned several snippets of conversation which could also be integrated in the EFL classroom with hardly any need to "doctor the text". One of these is re-printed below:

(251) **The point is**, we're both trying to teach you the same thing, to be a winner, not a Rosie Ruiz.
A Rosie Who-now?
Okay, Goldfarb.
One last lesson before I go... That lesson was about Rosie Ruiz, a world-class runner in the' 80s, famous for winning the Boston Marathon by **taking the subway**. […]
This woman cheated to win the Boston Marathon?
And no one noticed?
Nope. Everybody **was too caught up in the excitement** that an unfit woman who knew nothing about the sport didn't even **break a sweat** while **shattering a world record**.
Wow. That is so wrong.
Yeah.
She **took the subway**?
The subway!

[61] On english-corpora.org, words can be capitalised to search for all forms of a lemma (i.e., here *medal* and *medals*).

Everyone has a **sports hero**, and Rosie Ruiz is mine.
Controversial, but okay.
That's why during the mile run, I'll jump into your car, and you can drive me to the **finish line**?
Aw, kiddo. I... I can't help you.
Cause cheating is wrong?
No!
Cheating's that rush that **keeps me ticking**.
I can't help you 'cause I don't have a license.
<COCA: The Goldbergs (2019)>

At this stage, it is worth noting that drawing on conversational language from films and TV series would not necessarily lead to a lexico-grammatical impoverishment of the language input students are exposed to – on the contrary. Excerpt (251), for example, features a number of idiomatic collocations to which teachers would do well to draw learners' attention, e.g., *take the subway, be caught up in the excitement, break a sweat, shatter a world record,* etc. It is worth noting that in contrast to the vocabulary conveyed in excerpt (248) which consists foremost of individual nouns, excerpt (251) features many high-frequency verb + noun collocations (see also Barlow 2003: 7). This is crucial as much research has shown that the acquisition of such patterns of co-occurrence (e.g., collocations, chunks, lexical bundles, etc.) is essential to developing both fluency and accuracy in a foreign language (see, among others, Altenberg & Granger 2001; Cowie 1998; Herbst 1996; Hoey 2005; Hunston & Francis 2000; Langacker 2001; Nesselhauf 2005). Following a usage-based L2 instruction paradigm (see 1.4), these patterns of co-occurrence are conceptualised as constructions. The comparison of excerpts (248) and (251) illustrates the scarcity of constructions in stereotypically textbook-like texts. This is highly problematic because, as eloquently put by Herbst (2016: 77):

> If "it's constructions all the way down" (Goldberg 2006: 18) and language learning consists of the learning of constructions, then language teaching should consist of the teaching of constructions.

Unfortunately, however, the pervasiveness of constructions in language has yet to be fully grasped by those involved in L2 teaching (Ellis, Römer & O'Donnell 2016; Pérez-Paredes, Mark & O'Keeffe 2020; Tyler, Ortega & Uno 2018). In particular, excerpts such as (248) suggest that textbook authors seemingly continue to be tasked with the artificial insertion of countless individual nouns in textbook representations of spontaneous, spoken English, e.g., in (248), *strength, courage, determination, generosity, hope*, etc., at the detriment of frequent and idiomatic collocations.

Returning to example (251), it can also be hypothesised that, from a motivational point of view, students are more likely to learn and remember the term *driving license* if they first came across it in a text such as (251) in which there is a genuine communicative need to understand that, here, *license* refers to *driving license* to make sense of the conversation. More generally, it is fair to say that students are more likely to be intrinsically motivated when asked to engage with materials that they,

themselves, recognise as "authentic" and "real" (Gilmore 2011). In advocating for the inclusion of corpus-informed spoken grammar in ELT, Carter & McCarthy (1996: 370) question whether deciding that learners need not be exposed to certain kinds of natural English might "not ignore a psychological reality in that all of us as language learners and teachers are intrigued by real discourse and by what native speakers do with it". The authors go on to convincingly argue that only offering contrived, simplified models of English in ELT materials amounts to holding back information which ultimately disempowers learners. Given that a large body of (corpus) linguistic research has now demonstrated and extensively documented pedagogically relevant lexico-grammatical differences between different registers of English (e.g., Biber et al. 2021; Carter 2014; Carter & McCarthy 2017; Carter & McCarthy 2006b), continuing to offer learners models of conversational English that are evidently based on written norms is no longer tenable.

Although calls to rely more on authentic data and corpus tools in ELT materials design now go back two or more decades (e.g., Conrad 2000; McEnery & Xiao 2011; Mindt 1996; Prowse 1998; Römer 2004a; 2006a; Sinclair 1991; 2004, to mention but a few), up until recently, textbook authors could be forgiven for lacking the skills and knowledge to source and/or adapt authentic materials. Today, however, the availability of a wealth of suitable online resources and ease of use of free corpus tools (to mention but a few: cqpweb.lancs.ac.uk/, corpora.lancs.ac.uk/bnclab/, english-corpora.org, yohasebe.com/tcse/, see Le Foll 2021c: Appendix for many more) considerably facilitate the task. In addition to spoken corpora and film/TV series, podcasts and videocasts, televised talk shows, radio discussions and interviews can also make for suitable sources of natural spoken English. An example from a broadcast discussion on the topic of sports heroes is printed as (252). It comes from a radio interview and shows that, whilst many aspects of interaction can only be meaningfully conveyed in video material, authentic audio materials can also be meaningfully integrated in the EFL classroom. Their transcripts typically lend themselves better to be printed as textbook dialogues (e.g., (252)) than those of spoken conversational corpora like the Spoken BNC2014 (e.g., (249)).

(252) MARTIN: **Well**, Jimi, what do you think?
Mr-IZRAEL: I think it's the end of the baseball hero. **I mean**, **it's been coming for some time**. But I think, **again**, **you know**, as L. Spence says, **you know**, this is kind of **a nail in the coffin**. **I mean, you know**, our kids should be **looking up to sports heroes anyway** but now we know that they can for sure, **you know**. And Howard, **you know**, I was curious, do you think he should **be in the Hall of Fame** after this?
Mr-WITT: No, I don't see how. **I mean, you know**. Michel said **the question comes down to whether** he knew he was taking **a banned substance**. The question is what was he taking anything for, **you know? I mean**, how are you supposed to explain to your kids, **well**, he might have been taking something but it was, **you know**, might have been okay. **I mean**, that's **shades of gray** there that just shouldn't even exist, **you know**, taken

```
after...
Dr-SPENCE: You take pills for headaches, don't you, Howard?
Mr-WITT: What's that?
Dr-SPENCE: Don't you take pills for headaches?
Mr-WITT: Well, I do take pills for headaches.
Mr-IFTIKAR: Yeah, you don't take pills to make your head grow, though.
<COCA: Tell Me More (NPR, 2007)>
```

Note that, in addition to an abundance of discourse markers (*well, I mean, you know, anyway*), excerpt (252) also features a number of frequent idioms and chunks with high communicative value (e.g., *it's been coming for some time, the final nail in the coffin, the question comes down to whether*, etc.). Going beyond its pure linguistic value, a text such as (252) also has the potential to trigger genuinely meaningful discussions among students, e.g., on who deserves to be considered a sports hero, whether athletes are role models for young people, or how doping is influencing professional sport, etc.

As discussed in 7.3.2.1, textbook conversations not only tend to display neat and predictable turn-taking with no hesitations or misunderstandings of any kind, they almost exclusively consist of "referential discourse" (Blyth 2009: 196). Their primary function is "transactional" (Gilmore 2007: 102) or "informational-cognitive" (Blyth 2003: 63, 68). As such, they overwhelmingly neglect the "psychosocial functions of language, such as the creation of solidarity or the display of aggression" (Blyth 2009: 196; see also Cook 2000). This means that pedagogical materials largely fail to represent the more interactional, relationship-building, or psychosocial functions of conversations that may involve "controversial and imaginary content, or emotionally charged interaction" (Cook 2000: 158). Unsurprisingly, it is exactly these kinds of situations that instructed L2 learners often struggle to navigate (Gilmore 2007). This observation ties in with the well-known fact that commercial constraints often lead textbook publishers to avoid contentious topics. This is particularly true of the global EFL/ESL textbook market: textbook authors are often explicitly required to abstain from any mention of "PARSNIP topics" (Politics, Alcohol, Religion, Sex, Narcotics, -*Isms*, and Pork; Gray 2010: 119; see also Dinh & Siregar 2021; Smith 2020: 21–22), thus contributing to the kind of bland and banal textbook dialogues that are typically associated with EFL textbooks (see also 4.3.2.3). Though European school textbook publishers face slightly different constraints, textbook dialogues depicting difficult relationships are few and far between (see, however, excerpt (212) for an exception).

Some may counter that authentic listening materials – e.g., those drawn from film, TV or radio as discussed above, as well as from podcasts or social media such as YouTube – are inappropriate for lower secondary school EFL teaching because natural delivery rates are too fast for non-proficient speakers of English. Indeed, this is likely one of the reasons why many textbook publishers prefer to feature scripted dialogues which are then performed by professional actors at prescribed delivery rates

deemed more appropriate for beginners. This belief, however, is not supported by the conclusions of empirical research on the effect of delivery rates on EFL learners' listening comprehension: it has repeatedly been shown that lower-than-average speech rates are not beneficiary to or, indeed, preferred by language learners (e.g., Blau 1990; Derwing 1990; 2001; Derwing et al. 2012; Griffiths 1990; 1991; Munro & Derwing 1998; Révész & Brunfaut 2013).

In spite of all the aforementioned advantages of using non-scripted conversational materials, in some cases, it might not be feasible or practical to source suitable, authentic spoken materials. In such cases, textbook authors would do well to draw on corpus data or, at least, the findings of corpus research to arrive at a more realistic portrayal of conversational English. For instance, unnatural-sounding excerpts could be improved by consulting a corpus such as the Spoken BNC2014 (freely accessible on cqpweb.lancs.ac.uk) and adding some of the frequent lexico-grammatical features of spontaneous, interactional speech with high loadings on the 'Spontaneous interactional vs. Edited informational' dimension. This is illustrated in excerpt (253) which is a revised version of the textbook dialogue printed at the beginning of this section as (248).

(253) Journalist: **I'm** Sally Gordon, reporting from Leicester Square in London and the place **is** full of sports fans. **Let's** see **who we** can talk **to**. **Excuse me**, Sir. Can **I** ask **you who's your** sports hero**?**
Dwayne: **Erm**, for **me**, **it'd definitely** have to be Chris Hoy, **you know**, the British track cyclist **who** won two gold medals. **I think [THATD]** he **really stands** for strength and and **I really admire** his courage **because**, **well**, he **just** never **gives** up.
Journalist: **Sure**. **And erm what** about **you**? **Who would you** say **is your** national hero**?**
Donna: **Erm**, **actually**, **I'm** American **so** Kobe Bryant, **for sure**. **We're kind of** very patriotic, **especially** when **it** comes to sports, **if you know what I mean**.
Journalist: **And would you** say **[THATD]** basketball is **your** sport **then**?
Donna: **Yeah I am** into basketball and that and, **you know**, **I think [THATD]** he**'s really** shown the world **we're** still the best at **it**!
Journalist: **Mm**.
Donna: **Oh** and **I** should**n't forget** Michael Phelps**, of course**.
Journalist: **Uhu**. **What makes you say that?**
Donna: **You kidding? I mean**, he**'s just** won **like** four gold medals and two silver.
Journalist: **Right**, he did, **didn't he?**
Donna: And he**'s** a record holder! **I guess** what **I'm saying is** the the dream came true.
Journalist: **Right**.
Donna: **Yeah**, he**'s just** incredible. **I mean that's** why **we** call him "the Baltimore Bullet" **because** he**'s all** about determination, generosity, hope... he**'s all** about **all these really** great values. **You see**, he**'s** he**'s** a role model! And **we'll** never **forget** him, **that's for sure**.

This revised version features more 'mental' verbs (e.g. THINK, FORGET), *that* omissions (marked [THATD]), contracted and negated verbs, present tense verbs, first- and second-person references, emphatics (*definitely*, *really*), causative subordination (*because)*, discourse and pragmatic markers (*well, you know, if you know what I mean, what I'm saying is*), hedges (*kind of*), fillers and interjections (*erm, oh, yeah*), the modal *would*, demonstrative pronouns, and 'stranded' prepositions (e.g., *let's see who can talk to*) than the original textbook dialogue (248). As example (253) shows, such additions will also naturally lead to revised dialogues with lower type/token ratios, shorter average word lengths and, in particular, lower noun/verb ratios, which all contribute to higher scores on the first 'Spontaneous interactional vs. Edited informational' dimension (see 7.3.2), too.

Though it was not the object of this study to investigate the pedagogical efficacy of the language of school EFL textbooks, there is reason to believe that textbook dialogues with high Dimension 1 scores are better models for EFL learners to acquire the necessary skills to navigate real-life conversational situations (O'Keeffe, McCarthy & Carter 2007: 21). In particular, this includes the competent use of a variety of fluency-enhancing strategies to overcome planning phases and manage turn-taking. Interestingly, learner corpus research has shown that EFL learners significantly underuse discourse and vagueness markers as compared to native speakers and tend to rely more on filled and unfilled pauses and/or on a very limited set of such markers, instead (e.g., Müller 2005; Götz 2013; Gilquin 2016b; Dumont 2018). Wolk, Götz & Jäschke (2021: 4) have suggested that this frequently observed underuse of discourse markers in learner speech "might stem from the fact that an explicit teaching of discourse markers as a fluency-enhancing strategy has not been systematically integrated into EFL textbooks" (see also Gilquin 2016b). Though these studies were conducted on diverse learner populations who will have learnt with a variety of textbook and non-textbook materials, the results of the present study nevertheless lend support to this hypothesis – especially given that, in this respect, Textbook English is relatively homogenous: all nine textbook series of the TEC were found to, on average, misrepresent natural conversation in very similar ways.

8.2.2 Improving representations of informative texts

Although the results of Chapters 6 and 7 showed that, on the whole, the informative texts of the more advanced textbooks of the TEC are linguistically very similar to informative texts aimed at teenagers, some texts stood out as being more prototypically "textbook-like" than representative of this register. An example of such a text was already presented in 8.1: excerpt (245), about soap operas, was written in the style of a teenager magazine and, indeed, close inspection of the random effects associated with the Teen Vogue subcorpus of the Info Teens corpus (see 3.3.2.3) in 7.3.2 suggests that, on both the 'Spontaneous interactional vs. Edited informational' and the 'Narrative vs. Non-narrative' dimensions of the second PCA-based model,

Teen Vogue texts are closer to Textbook Informative texts than the rest of the reference Info Teens corpus of informative texts targeted at English-speaking teenagers.

Whilst this type of informative writing certainly has its place in secondary school EFL textbooks, some of the informative texts that also score high on the 'Spontaneous interactional vs. Edited informational' dimension make for rather unlikely candidate articles in such publication outlets. Text (254), for instance, is an informative text from a French textbook used in the fourth year of secondary English tuition that scores considerably higher than most texts of the reference Info Teens corpus on both the 'Spontaneous interactional vs. Edited informational' and the 'Pedagogically adapted vs. Natural' dimensions.

(254) Iwokrama, in what is called the Guiana shield, is a tropical rainforest reserve. **Because** there are only three other rainforest ecosystems like this in the world, Iwokrama is invaluable. It**'s a part of** "the lungs of the earth".

Moreover, it**'s** in a pristine state: it is **as if it had been untouched by humans**. **As though nobody had ever cut a tree**! But **indigenous people have lived there**: they have just done **so very** discreetly, leaving their natural environment **pretty much** intact.

Guyana's landscapes and wildlife are not only protected, they are also stunning: the Kaieteur Falls are majestic and it**'s** as if animals and plants were all "giant"! **You can** meet giant anteaters, giant water lilies, giant leaf frogs and giant otters!

Because this is such a unique place, Iwokrama has been made into an official reserve. The priority is the preservation of the rainforest. **But** this does not mean that Guyana refuses to make money out of the forest: it **just** has to be done sustainably so, with income for the communities that live there rather than gains for investors on the other side of the world. <TEC: Piece of Cake 3e>

The interactive web-based version of the textbook gives the impression that this is an authentic text and claims that it was "Adapted from Iwokrama.org." (https://www.lelivrescolaire.fr/page/16871655, 14.02.2022) – the official website of the *Iwokrama International Centre for Rain Forest Conservation and Development.* However, no text resembling the one featured in the textbook could be found on this informative website. In fact, text (254) has several tell-tale signs of a pedagogically doctored text. It hovers between different degrees of formality (e.g., *moreover* vs. *pretty much*) and, as such, sits rather uncomfortably between different registers. In this particular case, there is no doubt that the text was constructed around a pre-defined grammatical syllabus: the second paragraph features two *as if* conditional sentences (*as if it had been* and *as though nobody had ever*) and the textbook unit in which it is embedded includes several exercises on this grammatical structure (see

Fig. 88 and Fig. 89, both taken from (https://www.lelivrescolaire.fr/page/16871655, 14.02.2022). In both cases, the use of the past perfect in these two conditional sentences is clearly contrived: the present perfect would have sufficed. Moreover, whilst BE *untouched* is attested in corpora of naturally occurring English, the collocation REMAIN *untouched* is considerably more frequent and would have been a more idiomatic choice. It would also have helped address the fact that, across all registers, Textbook English features more occurrences of BE as a main verb per finite verb phrase (FVP) than in naturally occurring English.

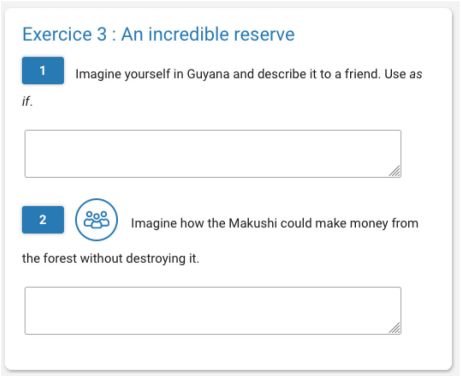


Fig. 88: Exercise featured below (254)

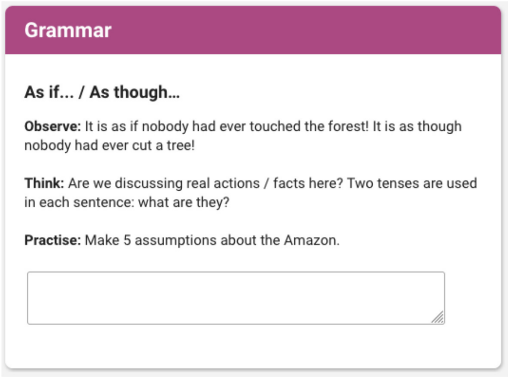


Fig. 89: Grammar box featured below (254)

Interestingly, a cursory look at the Iwokrama centre's website suffices to spot engaging materials which could be integrated with very few modifications into such a unit on the preservation of tropical rainforests for this particular proficiency level, e.g., (255). As an aside, it is also worth noting that excerpt (254) is potentially misleading in that the phrase "*have lived there*" without a temporal marker implies that indigenous people no longer live in the Iwokrama Forest. The reader will notice that excerpt (255) does not involve such ambiguity.

(255) DID YOU KNOW?
The Iwokrama Forest is located in central Guyana, approximately 300 km south of Georgetown, the capital. The area encompasses about 371,000 hectares and is covered in **lush, intact lowland tropical** forest. The wide range of intact habitats in the Iwokrama Forest supports a diverse **flora and fauna** with an estimated 1,500-2,000 higher plant species, 420 species of fish, 150 species of snakes, lizards and **frogs**, 500 species of birds and 180 species of mammals. […]

VIEWING TIPS!
Most mammals are secretive and can be hard to see. Since many mammals are nocturnal, a good way to see them is at night with the help of a headlamp. Fruiting trees are also a good place to see mammals as they congregate to feed. And always keep an eye on the ground for signs - especially tracks in the wet mud on the edge of pools. […]

DID YOU KNOW?
The Iwokrama Forest is in the homeland of the Makushi people, **who have lived in and used the forest for thousands of years**. People are a critical part of the **ecosystem** and the success of Iwokrama relies on the

combined skills of specialists and its community partners. <https://iwokrama.org/wp-content/uploads/2018/01/Iwokrama-Mammal-Guide-2017-Web.pdf, 14.02.2022>

Drawing on real-world resources to source ELT materials such as (255) affords learners valuable opportunities to acquire English in naturally occurring contexts. In a usage-based L2 instruction paradigm, context is known to be paramount to supporting language comprehension and pattern abstraction; however, it can also "be a stressor that introduces noise, complexity, and cognitive overload" (Tyler & Ortega 2018: 17–18). Hence, in some cases, it may make pedagogical sense to adapt authentic texts for specific proficiency levels and/or learner groups. Should textbook authors or teachers be worried that a text such as (255), drawn from a resource not especially targeted at L2 learners, could feature vocabulary that may be too demanding for their target learner group, user-friendly corpus-based tools can be used to identify potentially problematic lexical items. For instance, the free text analysis tool from english-corpora.org can be used to highlight the least frequent words based on frequency data from the COCA (Fig. 90).

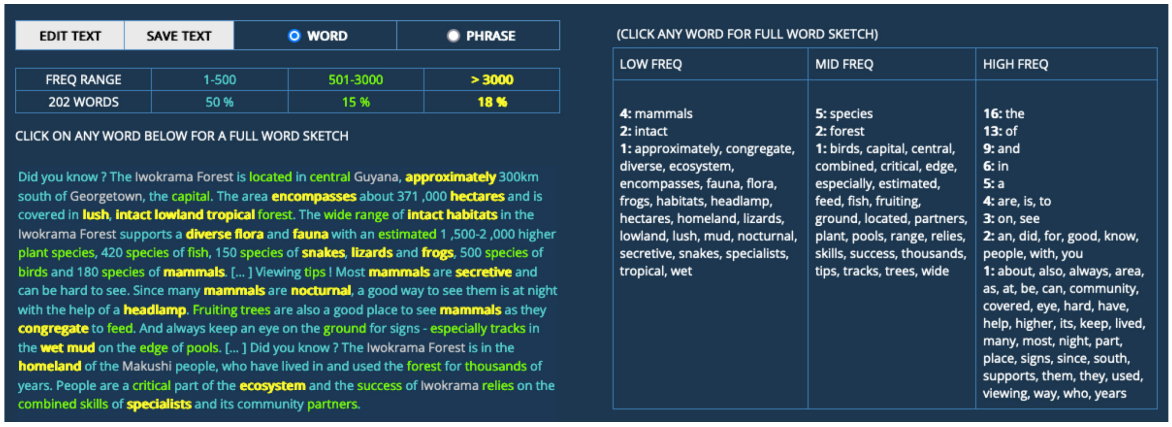


Fig. 90: Word frequency analysis with english-corpora.org (on the basis of COCA data) of excerpt (255)

Some of these low-frequency words (highlighted in blue and listed in the left-hand column of the table in Fig. 90) were already featured in the textbook excerpt (254) (e.g., *ecosystem, frog*). Others could easily be replaced by more frequent alternatives without compromising on the style of writing (e.g., *flora and fauna* → *plants and animals*). In choosing which low-frequency words to potentially replace, teachers and textbook authors would do well to focus on isotopy, i.e., on the lexical items that involve semantic redundance, rather than on those that involve strong collocational associations or make important contributions to a text's overall coherence (Hausmann 2005). For example, consider *lush, intact* and *tropical* in the first paragraph of excerpt (255). These are three low-frequency and semantically closely related words that are used to describe the *forest*. They need not all be included. Alternatively, clicking on any of the coloured words in the text analysis tool illustrated in Fig. 92 redirects the user to that word's 'word sketch' page. Fig. 91 is an excerpt of the word sketch for

the word *lush*. It shows a list of topics associated with the word, its most strongly associated collocates and makes suggestions for (potential) synonyms – all derived from the data of the COCA. On the basis of this information, teachers and materials designers may decide to replace *lush* with a higher-frequency word that learners are expected to already be familiar with, e.g., *beautiful* or *green*. In addition, learners' previous knowledge can be drawn on to make pedagogically informed adjustments to authentic texts. In the context of secondary EFL instruction, this also means taking account of which lexical items are cognates in the learners' L1 or school language. Thus, given that excerpt (254) is from a French EFL textbook, the adjectives *intact* and *tropical* should not pose a problem and can therefore be left unmodified. By the same token, with French L1 speakers as the target readership, *fertile* (see Fig. 91) could be chosen as an alternative to the word *lush* in this particular context.

Fig. 91: Part of the 'word sketch' page of the word *lush* as generated on english-corpora.org/coca

Note that, as shown on Fig. 92, and perhaps contrary to teachers' expectations, it is not the case that text (254), crafted specifically for pedagogical purposes, contains fewer low-frequency words than the one taken verbatim from *Iwokrama.org*.

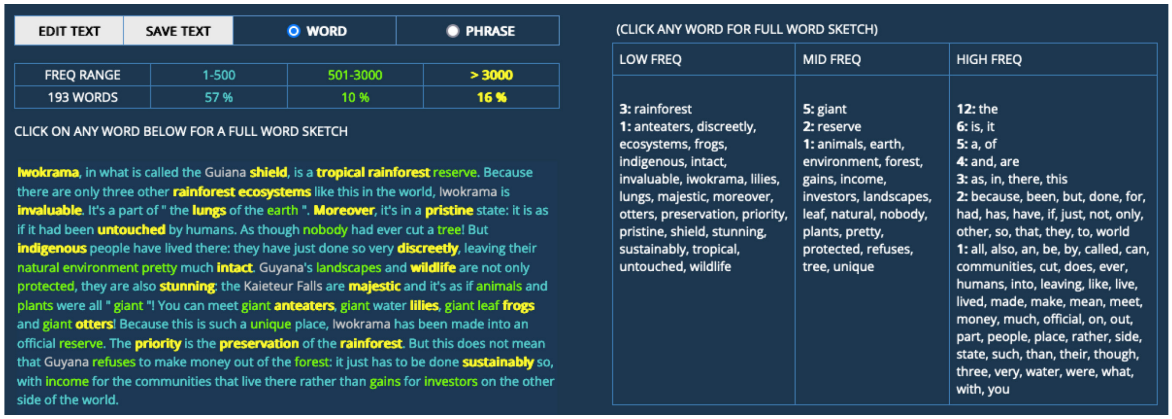


Fig. 92: Word frequency analysis with english-corpora.org (on the basis of COCA data) of excerpt (254)

Corpus tools can also be used to check whether a chosen alternative is suitable for any given register. For example, the 'word sketch' page of *moreover* (see Fig. 93 for an extract) shows that *moreover* is typical of academic writing but comparatively rare in news reports. Thus, in the context of (254), there is no doubt that the more versatile alternative *also* would have been more appropriate. As an aside, it is worth noting that *moreover* is known to be overused in Learner Englishes, in particular with French L1 speakers, including in registers where it is not the most idiomatic choice (see, e.g., Granger & Tribble 1998: 208).

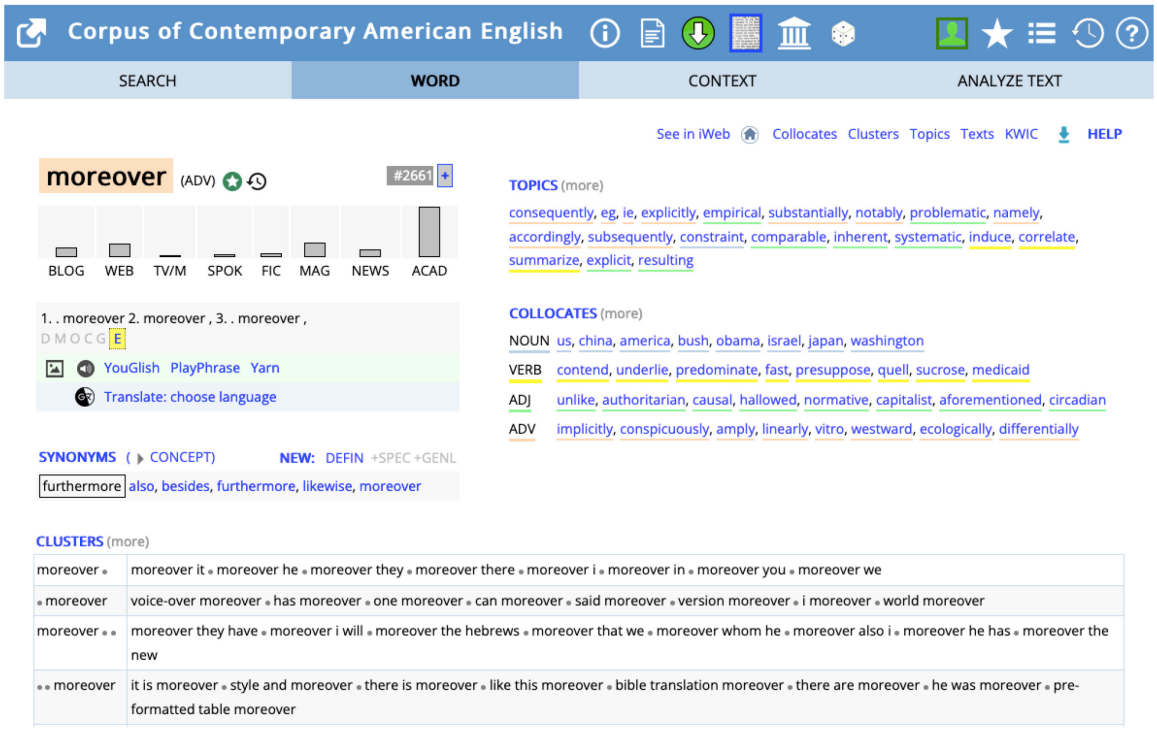


Fig. 93: Part of the "word sketch" page of the word *moreover* on english-corpora.org/coca

Compared with sourcing suitable conversational data for pedagogical use, finding suitable informative texts is much easier. For informative texts, good web searching strategies should suffice to find suitable texts on almost any topic of interest. A broad

range of different text registers are available on the internet. However, it is crucial that textbook authors, editors and teachers be aware of register-driven linguistic differences in order to avoid creating patchwork texts that result in unnatural sounding texts such as (254). Excerpts (256) and (257), which are both on the same topic yet clearly fulfil very different functions and are therefore written in very different styles, illustrate two such registers. Both could easily be adapted for use in the EFL classroom. Note how (256) would score lower than most textbook informative texts on the 'Spontaneous interactional vs. Edited informational' dimension as a result of its high frequency of passive constructions, coordinating conjunctions and non-finite *-ed* and *-ing* verb forms, among other features. It also boasts a number of useful collocations and constructions of the kind often missing from pedagogically adapted textbook texts (e.g., *SEEK to do sth., BOAST a wide range of, MAKE sth. an ideal sth., over the years, the likes of*, etc.).

(256) The Iwokrama International Centre for Rainforest Conservation is an autonomous, non-profit institution which **was set up** to manage the Iwokrama forest, **as** a "living laboratory".
The aim of the centre is to show how tropical forests can **be conserved** and sustainably **used** for ecological, social and economic benefits to local national and international communities. […]
The Iwokrama forest is in the homeland of the Makushi people who have lived and used the forest for thousands of years. **As such**, the Centre got its name Iwokrama from the range of mountains and **according to** the indigenous peoples, Iwokrama means 'place of refuge'.
Since its creation, Iwokrama has sought to **advance best practices** in the sustainable management of the world's remaining rain forests. It currently **boasts a wide range** of diverse flora and fauna **making it an ideal location** for bird-watching lovers, students, scientists, volunteers and interns **interested in seeing** and **experiencing** the untouched, lush rain forest. And, **over the years**, the Centre has attracted **the likes of** His Royal Highness Prince Charles; Prince Harry; President David Granger and First Lady Sandra Granger, Ministers of government, **among others**.
<https://guyanachronicle.com/2019/03/24/iwokrama-30-years-on, 14.02.2022>

As a marketing text with many imperatives, contracted verbs and first-person references, excerpt (257) would likely score higher than (256) on the 'Spontaneous interactional vs. Edited informational' dimension. It is very representative of its genre and, on the website from which it was sourced, is illustrated with many photos (e.g., of the *canopy walk*) that could support students' comprehension of the text without having to (systematically) resort to translation into students' first/school language.

(257) **If you're considering** travel to South America, **step outside the box** of typical Brazil **beach vacations** or Colombian coffee tours. Here, **we introduce you to** the beautiful country of Guyana, which will **feel like** an authentic slice of the "real" South America, from its **pristine rainforest** to its **welcoming villages**. Nature and wildlife lovers **are at home** here, where **first-hand exploration** of the **untrammeled countryside**

> is encouraged. There's an unmistakable pride in Guyana's people, as they **open their doors and hearts** to **curious travelers** seeking eco-friendly vacations, cultural immersion and a Lost World vibe. **Get** to Guyana now, before the crowds arrive. […]
> In a country that is 80 percent covered with **virgin tropical rainforest**, **it makes sense that** one of its top tourist attractions is a center focused on its conservation. **Feel as one with** the jungle as **you** tackle the canopy walk in the middle of the reserve - the birdwatching at this vantage point is unbelievable. The jaguar lives here - South America's largest cat - and while **we can't promise you a glimpse of** this elusive feline, **we** can almost guarantee that **you'll** meet ocelots, river turtles, otters, anteaters, caimans and more. **You** may even see a Goliath bird-eating spider as large as **your** fist! <https://navsumo.com/top-11-sights-to-see-in-guyana, 14.02.2022>

8.2.3 Towards a register approach to teaching EFL

In sum, the two preceding sections have demonstrated that, if the language to be taught in EFL textbooks is to be genuinely relevant to students' present and future lives, textbook materials must acknowledge that English is not a "monolithic block" (Rühlemann 2008: 681); but rather that, like all languages, English varies across different situational contexts of use. A large body of corpus linguistic research has demonstrated that various extralinguistic, socio-functional aspects of register have a direct impact on the linguistic features that characterise them (Crystal 2018: 490). Indeed, modern corpus-informed grammars of English no longer present "Standard English" as a single, homogenous language variety (as did, e.g., Quirk et al. 1985) but rather show how grammar varies across modes (e.g., in the *Cambridge Grammar of English*; Carter & McCarthy 2006b) and/or major registers (e.g., in the *Longman Grammar of Spoken and Written English*; Biber et al. 1999; 2021). This is necessary because vocabulary and grammar vary according to sociocultural, situational and functional contexts.

The model of intra-textbook variation presented in 7.3.1 has shown that a certain amount of situational/functional variation is already (implicitly) present across the texts featured in secondary school EFL textbooks. However, the continued prevalence of pedagogically adapted, artificial texts results in considerably less register-based linguistic variation than across situationally similar registers that EFL learners can be expected to encounter outside the classroom (see 7.3.2). In particular, most textbook representations of conversation remain very close to written norms and the grammatical phenomena that continue to form the backbone of textbook progressions are still taught as universally valid across all registers. Thus, it is fair to say that Conrad's (2000) optimistic prediction two decades ago that corpus linguistics could "revolutionise the teaching of grammar" and that, among other consequences, "[m]onolithic descriptions of English grammar [would] be replaced by register-specific descriptions" (Conrad 2000: 549) has not been fulfilled.

As the examples of the previous sections have shown, a register-sensitive approach to teaching English goes well beyond grammar. It involves all elements of the lexico-grammatical continuum. In fact, this study has repeatedly hinted at the fact that it also entails a re-appraisal of the semantic and pragmatic content of some textbook texts as modern school EFL materials largely continue to avoid potentially contentious topics (see also Timmis 2016: 5) and thus almost exclusively model harmonious, largely transactional interactions between proficient English users (see 8.2.1).

It is with similar concerns in mind that Rühlemann (2008) proposed a "register approach to teaching conversation". He convincingly argued that:

> the mismatch between school English and spoken English amounts to a mismatch between the end and the means deployed to reach it: SE-based [Standard English-based] school English fails to support learners in reaching their goal—to approximate to authentic English (Rühlemann 2008: 688).

The present study has shown that – although the gap between textbook dialogues and real-life English conversation is certainly the most worrying – this issue in fact concerns all registers. Adopting a register approach would certainly entail a number of long-term changes that cannot be expected to happen overnight. In many ways, however, it is quite surprising that register is still not firmly anchored in ELT, especially given that European school EFL curricula are now all aligned with the CEFR and that many of the *can-do* statements of the CEFR very much imply a register approach, e.g.:

> **B1** Can scan through straightforward**, factual texts in magazines**, **brochures** or on the **web**, identify what they are about and decide whether they contain information that might be of practical use. Can find and understand relevant information in everyday material, such as **letters**, **brochures** and **short official documents**. Can pick out important information about preparation and usage on the **labels on foodstuff** and **medicine**. Can assess whether an **article**, **report** or **review** is on the required topic. Can understand the important information in simple, clearly drafted **adverts** in **newspapers** or **magazines**, provided there are not too many abbreviations.
>
> **A2** Can find specific information in practical, concrete, predictable texts (e.g. **travel guidebooks, recipes**), provided they are produced in simple language. Can understand the main information in short and simple **descriptions of goods** in **brochures** and **websites** (e.g. portable digital devices, cameras). Can find specific, predictable information in simple everyday material such as **advertisements**, **prospectuses**, **menus**, **reference lists** and **timetables**. (CEFR "Reading for orientation"; Council of Europe 2020: 56 emphases added)

> **B2** Can read for pleasure with a large degree of independence, adapting style and speed of reading to different texts (e.g. **magazines**, more straightforward **novels**, **history books, biographies, travelogues, guides, lyrics, poems**), using appropriate reference sources selectively. Can read novels with a strong, narrative plot and that use straightforward, unelaborated language, provided they can take their time and use a dictionary.
>
> **B1** Can read **newspaper/magazine accounts of films, books, concerts**, etc. produced for a wider audience and understand the main points. Can understand simple **poems** and

> **song lyrics** provided these employ straightforward language and style. (CEFR "Reading as a leisure activity"; Council of Europe 2020: 59 emphases added)

This is true of all four categories of communicative language activities described in the CEFR: reception, production, interaction and mediation. For instance, for oral interaction, the CEFR (Council of Europe 2020: 71) mentions a range of linguistically quite distinct registers:

- interpersonal: "Conversation";
- evaluative: "Informal discussion (with friends)"; "Formal discussion (meetings)", "Goal-oriented collaboration";
- transactional: "Information exchange", "Obtaining goods and services", "Interviewing and being interviewed", and "Using telecommunications".

The present study has demonstrated that modern secondary school EFL textbooks very rarely model realistic "interpersonal" and "evaluative oral interactions" (see 8.2.4). Given the current state of affairs, we may therefore question how learners are supposed to achieve descriptors such as:

> **B2** Can establish a relationship with interlocutors through sympathetic questioning and expressions of agreement plus, if appropriate, comments about third parties or shared conditions. Can indicate reservations and reluctance, state conditions when agreeing to requests or granting permission, and ask for understanding of their own position. Can engage in extended conversation on most general topics in a clearly participatory fashion, even in a [audially/visually] noisy environment. Can sustain relationships with users of the target language without unintentionally amusing or irritating them or requiring them to behave other than they would with another proficient language user. Can convey degrees of emotion and highlight the personal significance of events and experiences. (Council of Europe 2020: 73)

Not only is a monolithic understanding of English not compatible with the CEFR and the many school curricula which are based on the framework, it is also not in line with a task-based language teaching (TBLT) approach (see Crawford & Zhang 2021). Indeed, in TBLT, learners are pushed to acquire language skills through real-world communicative situations which, as decades of corpus linguistic research have shown, will naturally have specific situational characteristics that, in turn, call for register-specific patterns of language use.

Decades after the so-called 'communicative turn' to foreign language teaching, it is somewhat disconcerting that learners are still not encouraged to communicate in differentiated ways depending on the situational context. In effect, students are left to deduce this by themselves. As exemplified by the interview excerpts in 8.2, many learners are aware that the kind of English they engage with outside the classroom is different from the kind of "Textbook English", "*Schulenglisch*" or "*inglés de libro*" that they learn at school.

We also noted in 1.6 that a large proportion of teenagers in Germany regularly engage with media in English (Feierabend et al. 2020: 48) and we can expect this trend to

be on the rise throughout Europe and beyond. What is striking is that, at least in Germany, there appears to be a genuine disconnect between students', sometimes quite extensive, contact with English outside the classroom and their English teachers' estimates of the quantity, quality and pedagogical value of that input (see, e.g., the results of a survey and focus group interviews with Year 9 students and their teachers in Grau 2009; for a more recent assessment of the situation and its underlying causes, see Blume 2020). A resolute commitment to a register approach could help to bridge this gap by helping learners and their teachers to understand the value of extracurricular English input whilst highlighting important linguistic differences between the various registers that learners encounter both *in* and *outside* the EFL classroom (see also Roberts & Cooke 2009). The author therefore proposes that a register approach can contribute to bringing together what Grau (2009: 161) refers to as "English from above" and "English from below", i.e., English as it is taught in the EFL classroom and English as it is "playfully taken up by German youths [or any other English L2 learners] and integrated into their own language" (Grau 2009: 163). In a similar vein, Willis' claim (2003: 224) that "[r]eal language provides a refreshing link between the classroom and the world outside" ties in with the need to do away with artificial, 'register-neutral' textbook language. In short, by adopting a register-sensitive approach, EFL teachers can acknowledge and validate learners' extracurricular exposure to English, whilst highlighting relevant register-driven differences in frequent, and therefore contextually appropriate, language use.

Although it is in line with the curricula of the educational systems examined in the present study (see 1.2) and the CEFR (see above), adopting a comprehensive register approach to secondary school EFL teaching undoubtedly entails a major overhaul in how English is taught. In particular, it carries far-reaching implications for EFL teachers and all those involved in pre- and in-service teacher education, as well as textbook authors, editors and publishers. Implications for teacher training will first be sketched out in 8.2.4. This will be followed by some thoughts as to how the textbook publishing industry could contribute to such a shift in perspective in 8.2.5.

8.2.4 Implications for teacher education

The introduction of a register approach to secondary EFL instruction can only succeed if a number of prerequisites are met. The first perhaps obvious – but by no means trivial – prerequisite is that (future) English teachers be aware of and knowledgeable about register-driven variation. A recent survey of 80 English schoolteachers from Sweden and Germany suggests that this is not yet common knowledge: in their answers on target language norms, many of the surveyed teachers appear to perceive standard target varieties of English as stable and homogenous entities (Forsberg, Mohr & Jansen 2019; Mohr, Jansen & Forsberg 2019). This is also the author's impression having worked with pre-service teachers of English at a German university for the past five years. Whilst student teachers are familiar with

some of the most obvious differences between regional (mostly inner-circle) varieties of English (especially lexical and pronunciation differences between British and American English), on the whole, they are largely oblivious to register-based linguistic variation. In fact, they frequently react with scepticism at any suggestion that lexico-grammar be subject to any form of situational variation (see also, more broadly, Wiese et al. 2017 regarding teachers' attitudes towards linguistic diversity in Germany; and Hall et al. 2017 for a report of similar teacher beliefs in the context of ELT in China). This does not come as a surprise since these pre-service teachers have themselves learnt English (and other foreign languages) in educational systems that do not foresee such variation.

Hence, the first implication for teacher education resides in the content of future EFL teachers' English linguistics classes which, at least in Germany, are usually limited to an introductory lecture followed by one or two elective seminars on more specialised topics (Jansen, Mohr & Forsberg 2021: 67–69). Bridges between linguistics as an academic discipline and student teachers' future careers are not always made explicit and many students fail to grasp the connection between what they are expected to learn in these courses and their future professional roles (see, e.g., Diehr 2018; 2020; Siepmann 2018b; Sommer 2020). In fact, in many cases, the practical English language classes they attend at university continue to propagate the myth of 'proper', 'Standard English' as a homogenous, register-neutral entity. Often, the only exception made is for Academic English – though, here too, it is not infrequent for the language of lectures, conference presentations and journal articles to be considered as one register and taught as such, resulting in some students confusing spoken and written academic discourse. Student teachers' university-level language practice classes therefore also have a crucial role to play in paving the way for a register-sensitive approach to EFL teaching.

At this stage, it is worth remembering that the development of register-sensitive language skills is a "universal process" that both L1 and L2 users develop over time (Gray & Egbert 2021: 177). Jansen et al. (2021) suggest that EFL student teachers may need to first reflect on the sociolinguistic implications of standard varieties in their L1(s) before they can begin to question Standard English ideologies and their implications in the EFL classroom. The same principle may also apply to register awareness: it may be beneficial to elucidate how language varies across different registers in students' L1 before transferring this sociolinguistic awareness to English.

Increased sociolinguistic knowledge, alone, however, will not suffice to bring about any meaningful change in the EFL classroom. The second, essential prerequisite to the successful introduction of a register approach in ELT entails a fundamental change in teachers' attitudes. This is something that should be addressed in pre-

service English/foreign language education classes. On the introduction of conversational grammar in the EFL classroom, Rühlemann (2008: 682) notes that:

> many teachers are likely to perceive the advent of conversational grammar as a threat to dearly held habits and convictions. To them, conversational grammar may simply be 'bad grammar' and, hence, not worth teaching.

In many cases, a register-sensitive approach will entail abandoning the conveniently simplistic dichotomy of 'correct' vs. 'incorrect' across all situations of use. Instead, it calls on notions of frequency and 'appropriateness' within specific contexts of use. Thus, rather than being able to apply a single rule:

> in a register approach, what is appropriate depends on the register and the specific set of conditions in that register constraining the use of the form in question (Rühlemann 2008: 682).

Resistance to long-held teacher beliefs and socially entrenched expectations that particular lexico-grammatical structures are either 'right' or 'wrong' is to be expected – though it remains to add that this is hardly a radically new idea, either (see, e.g., the concepts of 'appropriateness' and 'conformity' in traditional stylistics Crystal & Davy 1969: 4–7, 149–150).

Beyond gaining the necessary sociolinguistic knowledge and developing the willingness to apply this knowledge in the classroom, Rühlemann (2008: 682) emphasises that practical concerns will also need to be addressed:

> [t]he problem arising is less that correctness may be a dearly held notion that is hard to dispense with than rather that appropriateness is more difficult to handle.

This issue of added complexity is also acknowledged by Koch & Oesterreicher (2011: 276):

> Der Unterricht wird freilich durch die Berücksichtigung konzeptioneller Varianz für den Schüler keineswegs leichter. Dieser muss nunmehr alternative Regeln erlernen und in der Lage sein, sie situationsadäquat anzuwenden [However, by no means does taking contextual variation into account make learning any easier for students. They must now learn alternative rules and be able to apply them appropriately according to the situation].

The truth is, whether we like it or not, no language is monolithic. Corpus linguistics and, in particular, corpus-based register analysis has provided ample quantitative evidence supporting "the reality of underlying functional dimensions of language use" (Egbert & Biber 2018: 271). In fact, even young learners appreciate that the kind of English they engage with outside the classroom, e.g., on social media, is different from what they are expected to produce in an academic essay. In other words, rather than complicating teachers' jobs, a register approach can instead help foreign language teachers explain, on the one hand, why some structures may be "grammatically acceptable" yet not appropriate in specific contexts of use and, on the other, why other structures may be widely attested and therefore idiomatic in some registers, yet very rare and therefore inappropriate in other situational contexts. Of course, this is also true in learners' L1/school language; thus, it may be beneficial to first raise

awareness and illustrate the principle of register-specific patterns of language use in students' L1/school language before applying it to students' L2.
Ultimately, the author is convinced that if the foremost aim of school foreign language teaching is to develop learners' communicative competence, language learning cannot be detached from the situational contexts in which this communication is to take place. This is not to say that all aspects of the English lexico-grammatical system ought to be subdivided into an array of registers from the very first stages of language acquisition. Like all aspects of language acquisition, the process will necessarily be gradual. Learners will need to be encouraged, over time, to develop register awareness and to vary their language use according to different situational contexts. It probably makes sense to begin with key distinctions between the two poles of broad oral/informal/immediate vs. literate/formal/distant continuum of English variation (see also Chafe 1982; Koch & Oesterreicher 1985; 2011) before moving to finer-grained distinctions as learners develop their language skills and expand the repertoire of communicative situations they are expected to master in English.

As shown in 8.2.1 and 8.2.2, even basic corpus literacy can go a long way in helping teachers to source and adapt suitable teaching materials. The need to systematically integrate corpus literacy in the curricula of English teacher training study programmes is thus the third implication for teacher education (see also, among others, Callies 2016; 2019; Leńko-Szymańska 2017). Corpus linguistics can, to begin with, be used as an "eye-opener" for pre- and in-service teachers to understand why the doctrinal correct/incorrect dichotomy is not only unhelpful, but also often inaccurate in light of real language data. It is not within the scope of the present chapter to describe the kinds of corpus-based data-driven learning activities that may be used to introduce (student) teachers to register-based variation in English, but the evaluation of a project-based seminar conducted by the author (Le Foll 2020b; 2021d) suggests that activities that encourage students to debunk normative linguistic myths are particularly effective (see also the "surprise-the-teacher" modules suggested by Mukherjee 2004; as well as numerous books with suggestions for activities, e.g.: Bennett 2010; Crosthwaite 2020; Friginal 2018; Le Foll 2021c; O'Keeffe, McCarthy & Carter 2007; Pérez-Paredes & Mark 2021; Timmis 2015).

In this context, corpus literacy is to be understood as a subset of skills belonging, more broadly, to teachers' professional (critical) digital literacy and competence. Even pre-COVID-19, studies had shown that, whilst many teachers were interested in using more digital tools and media in their instruction, many were also acutely aware of their limited knowledge and competence in this area (e.g., Diz-Otero et al. 2022; Rohleder 2019). The pressing need for professional development opportunities in this domain was made all the more evident during the COVID-19 pandemic when teachers were forced to shift to online teaching with little to no preparation and, in many cases, suitable devices and/or infrastructure (see, e.g., Kerres 2020; Starkey et al.

2021; van de Werfhorst, Kessenich & Geven 2020). Reflecting on the situation in Germany, Blume (2020: 890) suggested that teachers often lack both the necessary theoretical and practical knowledge to develop their own materials and make meaningful use of web-based tools and materials. Studies suggest that the situation is comparable in France, Spain and across Europe (see, e.g., Fominykh et al. 2021). Few teachers appear to be aware of high-quality resources other than those proposed by the handful of textbook publishers that dominate each domestic market. Thus, in addition to raising awareness of what constitutes "Textbook English" from a linguistic point of view and reflecting, more generally, on the advantages and limitations of commercially published materials, teacher education in the 21st century also ought to focus on 'technological pedagogical content knowledge' (Mishra & Koehler 2006) and aim to develop future teachers' critical digital literacy and competence. Such courses would likely need to begin with relatively basic, general professional skills such as effective web searching strategies before moving to more complex, subject-specific competences such as ELT materials design and adaptation. The aforementioned pilot project led by the author demonstrated the potential and impact of project-based seminars in which student teachers engage in creating, adapting and reusing Open Educational Resources (OER) in ways that can create bridges not only between theory and practice, but also between pre-service teacher education, in-service teacher training and continuous professional development (Le Foll 2021c; see also Kosmas et al. 2021; Vyatkina 2020).

Although 8.2.1 and 8.2.2 explained how teachers can easily find and, if necessary, adapt authentic texts to create pedagogical materials for their students, the author is not suggesting that, at least at lower secondary school level, we do away with foreign language textbooks altogether. For a start, teachers' workloads in most European school systems simply do not allow enough time for this to be feasible. But even if time were not a constraint, the reality is that most teacher education programmes currently do not adequately prepare teachers for this task. In fact, it has been argued that commercially published textbooks – together with their handy multimodal packages consisting of texts, tasks and exercises, teacher handbooks, assessment materials, additional worksheets, games, etc. – play a crucial role in supporting inexperienced teachers. Schäfer (2003: 305) goes as far as to claim:

> Wer die Abschaffung des Lehrbuchs fordert, sollte gute Alternativen zu bieten haben. Immerhin bietet es Halteseile für den unsicheren Lehrer und schützt den Schüler vor dem schlechten Lehrer [Those who claim we should do away with textbooks ought to propose good alternative offers. After all, textbooks act as safety lines for insecure teachers and protect pupils from incompetent teachers].

The metaphor may seem like an exaggeration to some but, even if there is only a semblance of truth in the statement, it points to an alarming situation. For a start, it begs the question as to why teachers are placed in such a perilous situation that "safety lines" are necessary in the first place. It also places a disproportionate amount

of responsibility on the textbook industry that appears to be effectively tasked with filling glaring gaps in university teacher education with bite-size on-the-job training materials. Hence, a final yet crucial implication for teacher education consists in a much stronger emphasis, in pre- and in-service training, on the selection and use of pedagogical materials, including the considered and deliberate use of commercially-published textbook materials – many of which, it is worth reminding, are of excellent quality. At the end of the day, textbooks are not categorically 'good' or 'bad', 'suitable' or 'unsuitable'; however, the way they are exploited in the classroom may be effective or not. It is very much a case of: "Coursebooks don't kill learning, bad teachers kill learning" (Chong 2012). Thus, teacher education programmes would also do well to emphasise that:

> Kein Lehrwerk passt von selber zu jeder Lernsituation - und schon gar nicht gleichermaßen zu den Bedürfnissen eines jeden Lerners in einer größeren Lerngruppe: Lehrwerke sind vom Prinzip her auf aktive Interpretation angelegt [No textbook will, in and of itself, suit every learning situation – and can certainly not be expected to fulfil the needs of every learner in larger learning groups: textbooks are fundamentally designed to be actively interpreted]. (Vielau 2005: n.p.; see also, e.g., Nold 1998)

In sum, this section has made clear that teacher education – at all stages of teachers' careers – has a paramount role to play in addressing teachers' gaps in knowledge, competence, attitudes and beliefs on a range of issues relevant to successfully implementing a register approach in the secondary EFL classroom. These include raising awareness of what constitutes "Textbook English" and challenging the status of the textbook as "*the* authorised/legitimated educational medium for language learning" (Canale 2021: 1; emphasis original) on all language-related matters and as the best or "safest" way to implement the curriculum. It also involves developing teachers' own register awareness, both in their L1(s) and L2(s), by placing a stronger emphasis on the acquisition of sociolinguistic knowledge in pre-service teacher education. Furthermore, it entails a shift away from register-neutral dichotomies of *right* or *wrong* towards situationally-dependent notions of *frequency*, *idiomaticity* and *appropriateness* in specific contexts of use. Finally, this section has also pointed to the need for teacher education curricula to systematically integrate the building of both theoretical and practical skills in ELT materials design, including sourcing suitable authentic texts from online resources, adapting them to learners' needs, and making competent use of corpora and corpus tools.

8.2.5 Implications for materials design

As we have seen, at lower secondary school level, foreign language textbooks are seen as "indispensable" (Leroy 2012: 62) and, for a whole host of reasons, are unlikely to become obsolete any time soon. Hence, it goes without saying that this paradigm shift towards a register approach to ELT cannot happen without the involvement of the textbook publishing industry.

It is worth noting that, of the three major textbook registers compared to equivalent registers from outside the EFL classroom in 6.3.2 and 7.3.2, Textbook Fiction was, across all linguistic analyses, found to be the closest to its corresponding reference corpus. We concluded that this finding is not particularly surprising given that many of these fictional texts are excerpts of published novels and short stories. Those that are not, however, have been shown to be either practically indistinguishable from texts originally written as novels or short stories (thus, demonstrating that, on the whole, textbook authors appear to have an excellent command of this register) or have been convincingly written with low proficiency level learners in mind (e.g., using present-tense narration and low lexical diversity for beginner learners). However, the same cannot be said of representations of the informative and, in particular, conversation texts featured in the textbooks of the TEC. In contrast to the fictional texts typically found in EFL textbooks, the majority of these texts are crafted by textbook authors, presumably following strict pedagogically motivated guidelines.

Adopting a register approach would require textbook authors and editors to systematically account for register-driven linguistic variation when selecting, adapting and drafting textbook texts. How this could be achieved using corpus data and tools has already been exemplified in 8.2.1 and 8.2.2. These suggestions and recommendations are by no means new or particularly innovative. More than three decades ago, Sinclair (1991: 39–51) already explained why lexicographers and other applied linguists would do well to rely less on a combination of existing descriptions of languages and (native-speaker) introspection and more on attested language data in the form of large corpora. Yet, whilst it is true that corpora have since revolutionised the development of (learner) dictionaries and (many) reference grammars (see 1.8), textbook publishers have seemingly been much slower to follow this trend. That so few textbook publishers have latched onto the potential of corpus linguistics over the past three decades may seem particularly surprising given that corpora and corpus tools are more accessible than ever. However, Nelson (2022) notes that corpora are scarcely mentioned in reviews of ELT materials development such as Tomlinson (2008; 2012; 2013a) and Garton & Graves (2014; Graves & Garton 2019). By the same token, Meunier & Reppen's (2015: 501) encouraging report that "there has been a significant increase in corpus-informed teaching materials" cannot be confirmed for the European school EFL textbook industry. Whilst some of the large Anglo-Saxon publishing houses operating on the global (mostly adult) EFL/ESL market have invested in their own corpus resources and expertise and now advertise many of their products as "corpus-informed" (see 1.8), only one series of the TEC (*English in Mind for Spanish Speakers,* Cambridge University Press) explicitly states that it incorporates insights from (in this case, learner) corpus data:

> 'Get it right!' section based on information from the unique Cambridge Learner Corpus tackle problem areas common to learners of each level.
> (https://www.cambridge.org/us/cambridgeenglish/catalog/secondary/english-mind-2nd-edition, accessed 25 Jan. 2022)

At the turn of the millennium, Conrad (2000: 549) prophesised that corpus-based descriptions of language had "the potential to revolutionize the teaching of grammar" but, for this to happen, she highlighted the crucial role of materials designers (Conrad 2000: 557). As we have seen, in secondary school EFL materials design, the "corpus revolution" has yet to materialise in any significant way. As briefly outlined in 1.7, the school EFL textbook industry operates under very (national and regional) specific constraints and, as a result of these many constraints, is known to be particularly resistant to change.

In advocating for the use of corpora in materials design, there is sometimes a misconception that corpus linguists believe that frequency of use should override all other considerations. This assumption is misguided on several grounds. For a start, corpus linguistics necessarily involves a combination of quantitative and qualitative analyses (Bennett 2010: 7). Hence, when selecting the items to be included in pedagogical materials, textbook authors and editors cannot rely on frequency as the sole criterion. They will also need to consider additional factors such as salience, contingency, range, teachability, learnability, etc. (see, e.g., Ellis 2002; 2006; 2008). In other words, whilst corpora can provide valuable frequency-based information that is not accessible to (even native speaker) intuition, these quantitative statistical results need to be complemented with qualitative analyses. Making the same argument, McCarthy & Carter (2001: 338) assert that: "corpora can afford considerable benefits for classroom teaching, but the pedagogic process should be informed by the corpus, not driven or controlled by it." Nonetheless, the present author agrees with Biber & Conrad (2001: 335) that:

> [i]n the absence of other compelling factors (e.g., learnability at a given stage or basic knowledge required as a building block for later instruction), […] dramatic differences in frequency should be among the most important factors influencing pedagogical decisions.

In the context of secondary school textbooks designed for national markets, factors that are specific to certain L1s (i.e., 'learnability') can arguably best be teased out on the basis of learner corpus data (see, e.g., Granger 2015) and/or of contrastive analyses of L1 and L2 corpus data (see, e.g., Valero Garcés 1998). Seven of the nine textbook series in the TEC were specifically targeted at learners with a common L1/school language. Yet, echoing Granger's (2015: 494) observations that learner corpora's impact on textbooks has so far been more "more nominal than real", remarkably little contrastive metalinguistic information was provided by these textbooks. This seems like a missed opportunity since much research has confirmed that progress in L2 learning involves complex interactions between general language developmental processes and L1 constraints (e.g., Madlener 2018; Spada & Lightbown 1999). Corpus-based contrastive L1–L2 research and the findings of learner corpus research can provide textbook authors and editors with "information essential to producing customised syllabi applicable to teaching L2 learners of specific mother tongues" (Liu & Shaw 2001: 189). For instance, Winter & Le Foll (forthcoming)

sketch out ideas for such corpus-informed, register-sensitive and L1-specific "customised syllabi" with respect to the teaching of *if*-conditionals at lower secondary school level.

In sum, materials development can be both directly informed by the results of corpus queries formulated by the materials designers themselves and indirectly by incorporating the results of corpus linguistic research into the design process. Meunier & Reppen (2015: 501) list the following ways in which corpora can inform materials design:

- in helping select the linguistic target features (e.g. vocabulary, lexicogrammar; grammar);
- the amount of space in the text devoted to the features;
- in the sequencing of materials;
- through the inclusion of actual corpus data (e.g. lists of vocabulary or common lexico-grammar patterns);
- through the inclusion of information on register differences (e.g. conversation and academic prose);
- in the selection of the texts used in examples (e.g. do the texts accurately reflect the use of the target feature?).

Note that the penultimate bullet point refers specifically to the kind of register approach advocated for in 8.2.3, thus reiterating the major role that corpora and corpus tools can play in helping materials developers to promote register awareness in EFL teaching and learning.

Another idea, not touched upon in Meunier & Reppen's (2015: 501) summary, goes beyond the language featured in textbooks. It concerns the learning activities that textbooks propose. Though calls to incorporate corpus-based data-driven learning activities in the EFL/ESL classroom date back to the 1980s (e.g., Johns 1986), in secondary school contexts in particular, they remain an absolute exception (see, e.g., Chambers 2019; Boulton & Vyatkina 2020). The norm, as we have seen in 1.7, is for lower secondary school EFL teachers to largely follow the structure and activities of the textbook and since very few publishing houses have yet dared to include hands-on data-driven learning activities in their materials, only very few students (presumably those with particularly dedicated teachers who have attended at least one university seminar or continuous professional development course on corpus linguistics) benefit from these kinds of activities. Given that we know that learners already use an array of online resources to solve language issues, it would be wise for textbooks to include activities that guide students towards more trustworthy sources than they tend to choose (including, but not limited to, web-based corpus tools), teach them efficient and effective querying methods, as well as the necessary interpretative skills to make the best use of these resources (see Gilquin & Laporte 2021; Le Foll 2018b).

8.3 Limitations and methodological implications

In addition to the pedagogical implications discussed above, a number of methodological implications also arise from the present study. Each of the four analysis chapters (Ch. 4–7) concluded with a discussion of the strengths and limitations of the diverse corpus-linguistic methods employed. This section brings the most relevant aspects together to outline the present study's key methodological implications.

As with most corpus-based studies, arguably the most fundamental limitation concerns the representativeness of the corpus data from which conclusions were drawn and thus lies within the design of the corpora themselves. The decision-making processes that informed the design and compilation of both the TEC and the reference corpora were outlined in 3.3. As explained in 3.3.1.1, attempts were made to ensure that the selection of textbooks included in the TEC would be as representative as possible of the body of EFL textbooks currently[62] used at lower secondary school level in France, Germany and Spain. Similarly, considerable efforts were deployed to ensure that the three reference corpora used for comparisons with the Conversation, Fiction and Informative subcorpora of the TEC were carefully chosen/compiled to reflect as accurately as possible the kind of target language teenage EFL learners can be expected to aspire to interact with (see 3.3.2). However, given the lack of publicly available textbook sales figures and in light of other practical constraints, some convenience and/or arbitrary choices were undoubtedly made. As with all corpus studies, generalisability of the results beyond the corpus sample(s) analysed should not be assumed. For now, we cannot tell whether the results of the present study will generalise to other secondary school EFL textbooks used in France, Germany and Spain, let alone to textbooks designed for entirely different education systems and/or produced by very different textbook publishing cultures. That said, to the author's best knowledge, the TEC is the largest and most diverse corpus of contemporary secondary school EFL textbooks to date. It is also the only one to be entirely annotated for register. This is particularly important given that this study has demonstrated, across all four analysis chapters, that register is a major vector of language variation within textbooks and must therefore be accounted for when describing and/or evaluating Textbook English.

Another limitation common to most corpus-based studies resides in the fact that automated corpus queries are never 100% accurate. This being true, the present study reports on formal assessments of the accuracy of both the CQL queries used to query the corpora as tagged and lemmatised by Sketch Engine (in Chapters 4 and 5) and of the output of the MFTE (see 7.2.3). Moreover, inter-rater agreement scores were

[62] Or, rather, at the time of data collection in 2016–2017 (see 3.3.1.1).

computed as a means of assessing the reliability of the manual annotation schemes used in Chapters 3, 4 and 5.

As suggested in 7.4, the register annotation scheme of the TEC (outlined in 3.3.1.4) could be further developed – for instance by subdividing the Conversation category into 'private' and 'public' (e.g., broadcast) situational contexts. This would allow for additional comparisons with a corpus of TV and radio language for the 'public conversation' register, whilst the Spoken BNC2014 would only be used as a comparison benchmark for 'private conversation'. As a further improvement, the author would recommend that future endeavours to compile textbook corpora include additional XML tags that reflect the structure and layout of the textbooks. The simple mark-up scheme devised for the TEC only included a header with metadata on each textbook volume (see 3.3.1.3); however, provided there are enough resources for this additional manual annotation workload, it would also be worthwhile adding mark-up tags defining each textbook page, lesson unit and chapter. For the present study, this would have greatly facilitated the process of collating short textbook texts into meaningful units (e.g., in preparation for the MDAs, see 6.2.2). In future projects, it would allow for more fine-grained evaluations of textbooks' intended developmental linguistic progressions.

A further methodological takeaway message from the present study concerns the need to combine different (corpus-linguistic) methods to arrive at more robust answers. For instance, in Chapter 4, constructional analysis (see 4.2.3), correspondence analysis (see 4.2.4) and qualitative analyses of concordance lines and entire texts were combined to better grasp the similarities and differences in the verbs and semantic fields associated with the progressive in Textbook Conversation and Textbook Fiction as compared to the Spoken BNC2014 and the Youth Fiction corpus. A further example can be found in the case-study chapter of representations of the verb MAKE (Chapter 5) which demonstrated how one can arrive at different conclusions depending on the baseline used to calculate relative feature frequencies. The use of linguistically meaningful baselines was also one of the key features of the revised MDA framework outlined and implemented in Chapter 7 (see 7.2.4).

Methodological issues specific to comparisons of textbook language with that of naturally occurring registers have also been highlighted. Solutions to overcome issues related to the comparison of texts of vastly different lengths (see 6.2.2), the lack of punctuation in the transcriptions scheme of the Spoken BNC2014 (see 7.2.2) and the non-independence of texts from the same textbook series, web domain or novel (see 6.2.6) were discussed and implemented. The latter issue was solved by modelling these effects as random intercepts (and, whenever meaningful, also as random slopes, see 6.2.6) in mixed-effects models. In both Chapters 6 and 7, such models proved very useful to identify and quantify additional factors of linguistic variation, such as the

textbooks' target proficiency levels, series and countries of publication/use, and their interactions (see 6.3 and 7.3).

Both Chapters 6 and 7 confirmed that MDA (Biber 1988; 1995; Berber Sardinha & Veirano Pinto 2019) is a powerful framework with which Textbook English can successfully be described and compared to naturally occurring, 'real-life' English across several dimensions of linguistic variation. Using an unsupervised method meant that the texts of the TEC were not presumed to be linguistically different from those of the reference target language corpora: the models that emerged from these principal component analyses (PCAs), however, highlighted a number of pedagogically relevant discrepancies.

Whilst MDA is a dimension-reduction method that simplifies large correlation matrices of feature frequencies, it is also fair to say that, in terms of its implementation, it is a complex method. McEnery & Hardie (2011: 111) postulate that, in addition to replicability concerns, it is this complexity that "has inhibited the widespread uptake of what appears to be a useful technique". To a certain extent, both replicability and complexity concerns have been alleviated since Andrea Nini made the MAT (Nini 2014; 2019) available to the wider research community. As explained in 6.5, the GUI version of the MAT makes the method described in Chapter 6 accessible to researchers and materials developers with no coding experience. For this reason alone, additive MDA based on Biber's (1988) model of General Spoken and Written English, as implemented in Chapter 6, is worth considering for applied purposes such as textbook evaluation and/or preliminary linguistic explorations of Textbook English. That said, Section 6.5 also pointed to a number of limitations of additive MDA based on Biber (1988) for the study of school EFL textbooks, which led to the elaboration of the revised framework outlined in Chapter 7. In 7.4, it was argued that this framework is in fact easier to implement than the traditional Biberian framework. Furthermore, considerable efforts were deployed to make the method as accessible as possible by publishing the source code and full documentation for the tagger used in the analyses (the MFTE; Le Foll 2021a; 2021e), as well as the full analysis code for the PCA, its many visualisations and the statistical modelling of the dimension scores (see Online Appendix 7.2) – all of which were conducted in the open-source programming environment R (R Core Team 2020). As a result, it is hoped that the present study may serve as a springboard for future corpus-based research in textbook language description and evaluation and, more generally, for further methodological advances in multivariable corpus analysis.

The methodological framework outlined in Chapter 7 was shown to yield robust results that were successfully replicated over various (random and non-random) subsets of the data (see 7.4). As stated in 3.2.2, the full publication of the code and (copyright permitting) data used to run the analyses presented in this study paves

the way for additional, independent replications. Regrettably, this practice is currently far from the norm in (corpus) linguistics. It is therefore also hoped that this study may serve as a showcase for future studies by exemplifying how corpus-linguistic research can be made more transparent, reproducible and replicable.

8.4 Future research avenues

The present study is exploratory in nature. It opens many avenues for future research. Methodologically, it has shown how Textbook English may be examined across a broad range of different linguistic features both as a variety of English in its own right and in comparison to target reference varieties. It could be expanded by exploring the language of EFL textbooks and of other pedagogical materials, e.g., of online e-learning courses, used in different educational systems and at different levels.

Given that the present study has focused on frequent lexico-grammatical features, another avenue to be explored in future research focuses on the lexical input provided by EFL textbooks. For each textbook volume and series, the featured word and phraseme types can be extracted and their rates of repetition across each textbook volume and series can be calculated. The lexical input of the 42 textbook volumes and nine textbook series of the TEC may then be compared to examine the extent to which they present a common European core EFL lexical syllabus. In addition, the textbooks' lexical range can be compared to corpus-based lists such as the new General Service List (Brezina & Gablasova 2015) and the PHRASE List (Martinez & Schmitt 2012). Given the TEC's register annotation, it will also be possible to compare the phrasemes of the Conversation subcorpus of the TEC with the corpus-derived lists of the most frequent phrasemes in British and American spoken English (Fankhauser in preparation).

Needless to say, the revised MDA framework presented in 7.2 could also be applied to analyses of secondary school textbooks of other languages. It would be interesting to compare the present multi-dimensional models of Textbook English with those of other "textbook languages". Such comparisons might reveal that some of the observed characteristics of Textbook English are in fact universal features of what we might then call: "Textbook Language".

The present study has focused exclusively on the linguistic contents of EFL textbooks and can therefore only speculate as to their impact in and outside the EFL classroom. As vividly put by Cook (2002: 268),

> [i]t may be better to teach people how to draw with idealised squares and triangles than with idiosyncratic human faces. Or it may not. The job of applied linguists is to present evidence to demonstrate the learning basis for their claims […].

Indeed, and as made clear in 1.7, textbooks do not exist in a vacuum. Much research remains to be done on how teachers and students use textbooks in the classroom. Surprisingly few empirical studies have looked into how textbooks – i.e., not only their language, but also their structures, tasks and activities – mediate classroom interactions and learning outcomes (Rösler & Schart 2016: 490). Empirical data on the *status quo* in secondary EFL classrooms is urgently needed to understand the real impact of textbooks and subsequently develop research-informed pre- and in-service teacher training courses that address current problems and genuinely meet teachers' and learners' needs (see 8.2.4).

In addition to classroom-based investigations into textbook use, the results of the present study and follow-up corpus-based textbook language studies may be triangulated with findings from learner corpora to gain new insights into L2 learning processes. As early as 1998, McEnery and Kifle postulated that "[w]here textbooks are included in an exploration of L2 learning, they can explain differences between NS [native speaker] and NNS [non-native speaker] usage" (as cited in Tono 2004: 52). In this context, robust models of textbook language are potentially very useful because few large-scale research projects will realistically be able to investigate both the language of the textbooks that learners use and the language production of these same learners (though see Möller 2020 for such a research design). The hope is that, if the models of Textbook English presented in 7.3 are shown to be generalisable to further EFL textbooks, they may be used as a means of better understanding certain usage patterns that are more frequent in the language of instructed EFL learners than in that of naturalistic ESL learners (for first attempts in this direction, see Winter & Le Foll forthcoming on EFL learners' use of *if*-conditionals and; and Le Foll forthcoming on periphrastic causative constructions).

8.5 Conclusion

This study has provided a systematic empirical account of the relationship between English as it is presented to secondary school EFL learners and English outside the classroom. It has sought to provide what is likely, to date, the most comprehensive description of Textbook English as a distinct variety of English. Central to the thesis of this study is the notion of register. Indeed, throughout its diverse yet converging analyses, the present study has demonstrated that Textbook English cannot be adequately described without taking account of situationally determined linguistic variation. Whilst surprisingly few significant differences between the language of EFL textbooks used in secondary schools in France, Germany and Spain, or between that of the nine different textbook series of the TEC were observed, this study uncovered compelling interactions between text register and the target proficiency levels of the textbooks under study. The clusters of linguistic features responsible for these interactions were closely examined. Additionally, this study illustrated and explained

the key linguistic differences that define stereotypically textbook-like texts as opposed to situationally similar, 'real-life' texts.

Corroborating the findings of previous Textbook English studies, notably Mindt (1987; 1992; 1995b) and Römer (2004b; 2005), the present study identified a disconcerting gap between conversational English as it is presented in contemporary secondary school EFL textbooks and real-life conversation outside the classroom. In addition, it demonstrated that some representations of informative texts are potentially also problematic. Thus, the results of the present study reiterate a plea formulated decades earlier by, among others, Mindt (1996: 247) to: "bring textbooks for teaching English as a foreign language into closer correspondence with actual English". Since then, a large body of evidence from usage-based linguistic studies and related disciplines has consistently highlighted the strong connection between input exposure and L2 learners' developmental patterns (e.g., Achard & Niemeier 2004; Pérez-Paredes, Mark & O'Keeffe 2020; Tyler 2012; Tyler, Ortega & Uno 2018). Thus, more than ever, it makes sense to advocate for the design of "materials that reflect natural or authentic patterns of use" (Gurzynski-Weiss et al. 2018: 306) in order to improve the relevance and effectiveness of school EFL teaching materials.

Sections 8.2.1 and 8.2.2 illustrated concrete, practical ways in which authentic materials could be sourced and adapted for the EFL classroom with the use of freely available corpora and corpus tools. In 8.2.3, it has been argued that the use of materials that reflects naturally occurring English entails adopting a register approach to ELT that exposes learners to lexico-grammatical patterns of use in the form of situationally contextualised, meaningful constructions. Sections 8.2.4 and 8.2.5 spelt out the implications of such a register approach for teacher education and materials design. In addition to using a broad range of corpus-linguistic methods as descriptive and diagnostic tools in the four analysis chapters, this final chapter has underlined the essential role of corpus linguistics in guiding target language features, text choice and task design decisions in the development of new ELT materials.

In sum, the present pedagogically-driven corpus-based study has shown how corpus-linguistic methods can be used, on the one hand, to describe the language of textbooks from multiple perspectives and identify potential problems and, on the other, to propose solutions to improve future teaching materials by showing teachers and materials designers how freely available corpus resources and tools can be used to create and curate "meaningful content-rich contexts for language learning" (Pérez-Paredes, Mark & O'Keeffe 2020: 13). Hence, methodologically, the present study can be said to have "corpused" full circle.

References

Abello-Contesse, Christián & María Dolores López-Jiménez. 2010. The treatment of lexical collocations in EFL textbooks. In María Moreno Jaén, Fernando Serrano & María Calzada Pérez (eds.), *Exploring new paths in language pedagogy: lexis and corpus-based language teaching* (Equinox English Linguistics and ELT), 95–109. London; Oakville, CT: Equinox Pub.

Achard, Michel & Susanne Niemeier (eds.). 2004. *Cognitive linguistics, second language acquisition, and foreign language teaching* (Studies on Language Acquisition 18). Berlin: De Gruyter.

Ahmad, Hamdi & Robert McColl Millar. 2020. Review of Text Authenticity Relationship with Language Learner Motivation and Communicative Competence. *International Journal of English Linguistics* 10(5). 89. https://doi.org/10.5539/ijel.v10n5p89.

Aijmer, Karin. 2002. *English discourse particles: evidence from a corpus* (Studies in Corpus Linguistics 10). Amsterdam: John Benjamins.

Alejo González, Rafael, Ana Piquer Píriz & Guadalupe Reveriego Sierra. 2010. Phrasal verbs in EFL course books. In Sabine De Knop, Frank Boers & Antoon De Rycker (eds.), *Fostering Language Teaching Efficiency through Cognitive Linguistics* (Applications of Cognitive Linguistics 17), 59–78. Berlin: De Gruyter.

Alemi, Minoo & Ebrahim Isavi. 2012. Evaluation of interactional metadiscourse in EFL textbooks. *Advances in Asian Social Science* 2(1). 422–430.

Allaire, JJ, Kevin Ushey & Yuan Tang. 2019. *Reticulate: interface to "Python."* https://CRAN.R-project.org/package=reticulate.

Al-Surmi, Mansoor. 2012. Authenticity and TV Shows: A Multidimensional Analysis Perspective. *TESOL Quarterly* 671–694. https://doi.org/10.1002/tesq.33.

Altenberg, Bengt. 1989. Review of Douglas Biber (1988) Variation across speech and writing. *Studia Linguistica* 43(2). 167–174. https://doi.org/10.1111/j.1467-9582.1989.tb00800.x.

Altenberg, Bengt. 2002. Causative constructions in English and Swedish: A corpus-based contrastive study. In Bengt Altenberg & Sylviane Granger (eds.), *Lexis in contrast: corpus-based approaches* (Studies in Corpus Linguistics 7), 97–116. Amsterdam: John Benjamins.

Altenberg, Bengt & Sylviane Granger. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied linguistics* 22(2). 173–195.

Altosaar, Toomas, Louis ten Bosch, Guillaume Aimetti, Christos Koniaris, Kris Demuynck & Henk van den Heuvel. 2010. A Speech Corpus for Modeling Language Acquisition: CAREGIVER. In Calzolari, Nicoletta, Khalid Choukri, Bente Maeggard & Joseph Mariani (eds.),. Valetta, Malta: European Language Resources Association (ELRA).

Andersen, Roger W. & Yasuhiro Shirai. 1994. Discourse Motivations for Some Cognitive Acquisition Principles. *Studies in Second Language Acquisition* 16(2). 133–156. https://doi.org/10.1017/S0272263100012845.

Anton, Daniela. 2017. *Inter- und transkulturelles Lernen im Englischunterricht: eine didaktische Analyse einschlägiger Lehrbücher*. Heidelberg: Universitätsverlag Winter.

Archer, Dawn, Andrew Wilson & Paul Rayson. 2002. Introduction to the USAS category system. http://ucrel.lancs.ac.uk/usas/usas_guide.pdf (1 February, 2019).

Argamon, Shlomo. 2019. Register in computational language research. *Register Studies* 1(1). 100–135. https://doi.org/10.1075/rs.18015.arg.

Association for Computing Machinery, (ACM). 2020. Artifact Review and Badging version 1.1. https://www.acm.org/publications/policies/artifact-review-and-badging-current (21 August, 2021).

Atkins, Sue, Jeremy Clear & Nicholas Ostler. 1992. Corpus design criteria. *Literary and linguistic computing* 7(1). 1–16.

Baayen, R. Harald. 2012. Mixed-effects models. In Abigail C. Cohn, Cécile Fougeron & Marie K. Huffman (eds.), *The Oxford handbook of laboratory phonology* (Oxford Handbooks in Linguistics), 668–678. Oxford; New York: Oxford University Press.

Ballier, Nicolas, Ana Díaz Negrillo & Paul Thompson (eds.). 2013. *Automatic treatment and analysis of learner corpus data* (Studies in Corpus Linguistics volume 59). Amsterdam ; Philadelphia: John Benjamins Publishing Company.

Banerji, Nilanjana, Vineeta Gupta, Adam Kilgarriff & David Tugwell. 2013. Oxford Children's Corpus: A corpus of children's writing, reading, and education. *Corpus Linguistics* 315.

Barbieri, Federica. 2005. Quotative Use in American English: A Corpus-Based, Cross-Register Comparison. *Journal of English Linguistics* 33(3). 222–256. https://doi.org/10.1177/0075424205282667.

Barbieri, Federica & Suzanne EB Eckhardt. 2007. Applying Corpus-Based Findings to Form-Focused Instruction: The Case of Reported Speech. *Language Teaching Research* 11(3). 319–346. http://dx.doi.org/10.1177/1362168807077563.

Bardovi-Harlig, Kathleen. 2012. After process, then what? A longitudinal investigation of the progressive prototype in L2 English. In Emmanuelle Labeau & Inès Saddour (eds.), *Tense, aspect, and mood in first and second language acquisition* (Cahiers Chronos 24), 131–151. Amsterdam; New York: Rodopi.

Barlow, Michael. 2003. Corpora and language teaching. *KATE Forum* 27(1). 6–9.

Barlow, Michael & Suzanne Kemmer (eds.). 2000. *Usage-based models of language*. Stanford, Calif: CSLI Publications, Center for the Study of Language and Information.

Baroni, Marco & Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing* 21(3). 259–274. https://doi.org/10.1093/llc/fqi039.

Baroni, Marco & Stephanie Evert. 2009. Statistical methods for corpus exploitation. In Merja Kytö & Anke Lüdeling (eds.), *Corpus Linguistics: An International Handbook*, vol. 2, 777–803. Berlin: De Gruyter.

Baroni, Marco, Adam Kilgarriff, Jan Pomikálek & Pavel Rychlý. 2006. WebBootCaT: a web tool for instant corpora. In Elisa Corino, Carla Marello & Onesti Cristina (eds.), vol. 1, 123–132. Torino, Italy: European Association for Lexicography.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

Bausch, Karl-Richard (ed.). 2005. *Bildungsstandards für den Fremdsprachenunterricht auf dem Prüfstand: Arbeitspapiere der 25. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts* (Giessener Beiträge Zur Fremdsprachendidaktik). Tübingen: Gunter Narr Verlag.

Behrens, Heike. 2006. The input–output relationship in first language acquisition. *Language and cognitive processes* 21(1–3). 2–24.

Bell, Jan & Roger Gower. 2011. Writing course materials for the world: a great compromise. In Brian Tomlinson (ed.), *Materials Development in Language Teaching*, 135–150. Cambridge: Cambridge University Press.

Belli, Serap Atasever. 2017. An Analysis of Stative Verbs Used with the Progressive Aspect in Corpus-informed Textbooks. *English Language Teaching* 11(1). 120–135. https://doi.org/10.5539/elt.v11n1p120.

Belz, Anya, Shubham Agarwal, Anastasia Shimorina & Ehud Reiter. 2021. A Systematic Review of Reproducibility Research in Natural Language Processing. *arXiv:2103.07929 [cs]*. http://arxiv.org/abs/2103.07929 (26 July, 2021).

Bendix, Regina. 1997. *In Search of Authenticity: The Formation of Folklore Studies*. University of Wisconsin Press.

Bennett, Gena. 2010. *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. Ann Arbor, MI: University of Michigan Press. https://doi.org/10.3998/mpub.371534.

Berber Sardinha, Tony & Douglas Biber (eds.). 2014. *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber* (Studies in Corpus Linguistics (SCL) 60). Amsterdam: John Benjamins.

Berber Sardinha, Tony & Marcia Veirano Pinto. 2017. American television and off-screen registers: a corpus-based comparison. *Corpora* 12(1). 85–114. https://doi.org/10.3366/cor.2017.0110.

Berber Sardinha, Tony & Marcia Veirano Pinto (eds.). 2019. *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic. https://doi.org/10.5040/9781350023857.

Berber Sardinha, Tony, Marcia Veirano Pinto, Cristina Mayer, Maria Carolina Zuppardi & Carlos Henrique Kauffmann. 2019. Adding Registers to a Previous Multi-Dimensional Analysis. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 165–188. New York, NY: Bloomsbury.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18. https://doi.org/10.1515/ling-2017-0032.

Berns, Margie S., Kees De Bot & Uwe Hasebrink (eds.). 2007. *In the presence of English: Media and European youth* (Language Policy 7). New York: Springer.

Bezemer, Jeff & Gunther Kress. 2016. The Textbook in a Changing Multimodal Landscape. In Nina-Maria Klug & Hartmut Stöckl (eds.), *Handbuch Sprache im multimodalen Kontext*, 476–498. Berlin; Boston: De Gruyter.

Biber, Douglas. 1984. *A model of textual relations within the written and spoken modes*. University of Southern California PhD dissertation.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024.

Biber, Douglas. 1990. Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. *Literary and Linguistic Computing* 5(4). 257–269. https://doi.org/10.1093/llc/5.4.257.

Biber, Douglas. 1992. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes* 15(2). 133–163.

Biber, Douglas. 1993a. Representativeness in corpus design. *Literary and linguistic computing* 8(4). 243–257.

Biber, Douglas. 1993b. Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics* 19(2). 219–241.

Biber, Douglas. 1995. *Dimensions of Register Variation*. Cambridge, UK: Cambridge University Press.

Biber, Douglas. 2006. *University language: a corpus-based study of spoken and written registers* (Studies in Corpus Linguistics v. 23). Amsterdam: John Benjamins.

Biber, Douglas. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1). 9–37. https://doi.org/10.1515/cllt-2012-0002.

Biber, Douglas. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast* 14(1). 7–34. https://doi.org/10.1075/lic.14.1.02bib.

Biber, Douglas. 2019. Multi-Dimensional Analysis: A Historical Synopsis. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 11–26. London: Bloomsbury Academic. https://doi.org/10.5040/9781350023857.

Biber, Douglas. 2021. Corpus linguistics is for text-lovers. *Linguistics with a corpus*. https://linguisticswithacorpus.wordpress.com/2021/12/22/corpus-linguistics-is-for-text-lovers%EF%BF%BC/ (28 December, 2021).

Biber, Douglas & Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26(3). 263–286. https://doi.org/10.1016/j.esp.2006.08.003.

Biber, Douglas & Susan Conrad. 2001. Quantitative corpus-based research: Much more than bean counting. *TESOL quarterly* 35(2). 331–336.

Biber, Douglas & Susan Conrad. 2019. *Register, Genre, and Style*. Second edition. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd & Marie Helt. 2002. Speaking and Writing in the University: A Multidimensional Comparison. *TESOL Quarterly* 36(1). 9. https://doi.org/10.2307/3588359.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay & Alfredo Urzua. 2004. *Representing Language Use in the University: Analysis of the TOEFFL 2000 Spoken and Written Academic Language Corpus* (TOEFL Monograph Series). Princeton, NJ: Educational Testing Service.

Biber, Douglas & Jesse Egbert. 2016. Register Variation on the Searchable Web: A Multi-Dimensional Analysis. *Journal of English Linguistics* 44(2). 95–137. https://doi.org/10.1177/0075424216628955.

Biber, Douglas & Jesse Egbert. 2018. *Register Variation Online*. 1st edn. Cambridge University Press. https://doi.org/10.1017/9781316388228.

Biber, Douglas, Jesse Egbert, Bethany Gray, Rahel Oppliger & Benedikt Szmrecsanyi. 2016. Variationist versus text-linguistic approaches to grammatical change in English: nominal modifiers of head nouns. In Merja Kyto & Paivi Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*, 351–375. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139600231.022.

Biber, Douglas & Edward Finegan. 1994. Multi-dimensional analyses of authors' styles: Some case studies from the eighteenth century. In D. Ross & D. Brink (eds.), *Research in Humanities Computing*, vol. 3, 3–17. Oxford: Oxford University Press.

Biber, Douglas & Bethany Gray. 2013. Discourse characteristics of writing and speaking task types on the TOEFL IBT test: A lexico-grammatical analysis. *ETS Research Report Series* 2013(1). https://doi.org/10.1002/j.2333-8504.2013.tb02311.x.

Biber, Douglas & Bethany Gray. 2015. *Grammatical complexity in academic English: linguistic change in writing* (Studies in English Language). Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Biber, Douglas, Stig Johansson, Geoffrey N. Leech, Susan Conrad & Edward Finegan. 2021. *Grammar of Spoken and Written English*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.232.

Biber, Douglas & Randolph Quirk (eds.). 2012. *Longman Grammar of Spoken and Written English*. 10th edition. Harlow: Longman.

Biber, Douglas & Randi Reppen. 2002a. What Does Frequency Have to Do with Grammar Teaching? *Studies in Second Language Acquisition* 24(02). 199–208. https://doi.org/10.1017/S0272263102002048.

Biber, Douglas & Randi Reppen. 2002b. What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition* 24(02). https://doi.org/10.1017/S0272263102002048 (3 August, 2017).

Biebighäuser, Katrin, Marja Zibelius & Torben Schmidt. 2012. *Aufgaben 2.0: Konzepte, Materialien und Methoden für das Fremdsprachenlehren und -lernen mit digitalen Medien*. BoD – Books on Demand.

Bieswanger, Markus. 2012. Varieties of English in current English language teaching. *Stellenbosch Papers in Linguistics* 38(0). https://doi.org/10.5774/38-0-21.

Binnick, Robert I. 2020. Aspect and Aspectuality. In Bas Aarts, April McMahon & Lars Hinrichs (eds.), *The Handbook of English Linguistics*, 183–205. Wiley. https://doi.org/10.1002/9781119540618.ch11.

Birkland, Annie, Adeli Block, Justin Craft, Yourdanis Sedarous, Sky Wang, Wu Gou & Savithry Namboodiripad. 2022. Problematizing the "native speaker" in Linguistic Research: History of the term and ways forward. Presented at the Linguistic Society of America. https://osf.io/jufmg/ (16 January, 2022).

Bland, Susan Kesner. 1988. The Present Progressive in Discourse: Grammar versus Usage Revisited. *TESOL Quarterly* 22(1). 53–68. https://doi.org/10.2307/3587061.

Blau, Eileen K. 1990. The Effect of Syntax, Speed, and Pauses on Listening Comprehension. *TESOL Quarterly* 24(4). 746. https://doi.org/10.2307/3587129.

Blume, Carolyn. 2020. German Teachers' Digital Habitus and Their Pandemic Pedagogy. *Postdigital Science and Education* 2(3). 879–905. https://doi.org/10.1007/s42438-020-00174-9.

Blumenthal-Dramé, Alice. 2012. *Entrenchment in usage-based theories: what corpus data do and do not reveal about the mind* (Topics in English Linguistics 83). Berlin: Mouton De Gruyter.

Blyth, Carl. 2003. Playing games with literacy: the poetic function in the age of communicative language teaching. In Peter Charles Patrikis (ed.), *Reading between the lines: perspectives on foreign language literacy* (Yale Language Series), 60–73. New Haven: Yale University Press.

Blyth, Carl. 2009. From textbook to online materials: the changing ecology of foreign language publishing in the era of ICT. (Ed.) Michael J. Evans. *Foreign language learning with digital technology*. Bloomsbury Publishing 174–202.

Bohmann, Axel. 2017. *Variation in English world-wide: Varieties and genres in a quantitative perspective*. Austin: University of Texas PhD dissertation.

Bolker, Ben. 2020. General Linear Mixed Models FAQ. http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html (22 October, 2020).

Bot, Kees de, Wander Lowie & Marjolijn Verspoor. 2007. A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition* 10(1). 7–21. https://doi.org/10.1017/S1366728906002732.

Boulton, Alex & Tom Cobb. 2017. Corpus use in language learning: A meta-analysis. *Language Learning* 67(2). 348–393. https://doi.org/10.1111/lang.12224.

Boulton, Alex & Nina Vyatkina. 2020. Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning* 24.

Bragger, Jeannette D. & D. B. Rice. 2000. Foreign language materials: Yesterday, today, and tomorrow. In R. M. Terry (ed.), *Agents of change in a changing age*, 107–140. Lincolnwood, IL: National Textbook Company.

Brezina, V. & D. Gablasova. 2015. Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics* 36(1). 1–22. https://doi.org/10.1093/applin/amt018.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899.

Brezina, Vaclav, Abi Hawtin & Tony McEnery. 2021. The Written British National Corpus 2014 – design and comparability. *Text & Talk* 41(5–6). 595–615. https://doi.org/10.1515/text-2020-0052.

Brown, Dale. 2014. The power and authority of materials in the classroom ecology. *The Modern Language Journal* 98(2). 658–661.

Bruyn, Bert Le & Magali Paquot. 2021. *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press.

Burnard, Lou (ed.). 2007. Reference Guide for the British National Corpus (XML Edition). http://www.natcorp.ox.ac.uk/XMLedition/URG/ (17 September, 2021).

Burnard, Lou & Syd Bauman. 2021. TEI P5: Guidelines for electronic text encoding and interchange v. 4.3.0. TEI consortium. https://tei-c.org/Guidelines/P5/ (2 January, 2022).

Burton, Graham. 2012. Corpora and coursebooks: destined to be strangers forever? *Corpora* 7(1). https://doi.org/10.3366/corp.2012.0019.
Burton, Graham. 2020. Grammar. *ELT Journal* 74(2). 198–201. https://doi.org/10.1093/elt/ccaa004.
Burton, Graham Francis. 2019. *The canon of pedagogical grammar for ELT: a mixed methods study of its evolution, development and comparison with evidence on learner output*. Limerick: University of Limerick PhD thesis.
Bybee, Joan. 2010. *Language, Usage and Cognition*. Leiden: Cambridge University Press. http://www.myilibrary.com?id=263161 (11 February, 2020).
Bybee, Joan & Paul J. Hopper (eds.). 2001. *Frequency and the emergence of linguistic structure* (Typological Studies in Language 45). Amsterdam: John Benjamins.
Bybee, Joan L. 2007. *Frequency of use and the organization of language*. Oxford ; New York: Oxford University Press.
Callies, Marcus. 2016. Towards corpus literacy in language teacher education. 2.
Callies, Marcus. 2019. Integrating corpus literacy into language teacher education: The case of learner corpora. In Sandra Götz & Joybrato Mukherjee (eds.), *Learner Corpora and Language Teaching*, 245–263. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.201.12cal.
Cambridge University Press. 2017. Open Cambridge Learner English Corpus, available on Sketch Engine. https://www.sketchengine.eu/cambridge-learner-corpus/ (29 November, 2019).
Canale, Germán. 2021. The language textbook: representation, interaction & learning: conclusions. *Language, Culture and Curriculum* 34(2). 199–206. https://doi.org/10.1080/07908318.2020.1797081.
Candarli, Duygu. 2021. Linguistic characteristics of online academic forum posts across subregisters, L1 backgrounds, and grades. *Lingua* 103190. https://doi.org/10.1016/j.lingua.2021.103190.
Carrie, Erin. 2017. 'British is professional, American is urban': attitudes towards English reference accents in Spain. *International Journal of Applied Linguistics* 27(2). 427–447. https://doi.org/10.1111/ijal.12139.
Carter, R. & M. McCarthy. 1996. Correspondence. *ELT Journal* 50(4). 369–371. https://doi.org/10.1093/elt/50.4.369-b.
Carter, Ronald. 1998. Orders of reality: CANCODE, communication, and culture. *ELT Journal* 52. 1.
Carter, Ronald. 2014. Grammar and spoken English. In Caroline Coffin, Ann Hewings & Kieran O'Halloran (eds.), *Applying English Grammar: Corpus and Functional Approaches*, 25–39. Oxon: Routledge.
Carter, Ronald, Rebecca Hughes & Michael McCarthy. 1998. Telling tails: Grammar, the spoken language and materials development. In Brian Tomlinson (ed.), *Materials development in language teaching*, 67–86. Cambridge: Cambridge University Press.
Carter, Ronald & Michael McCarthy. 1995. Grammar and the Spoken Language. *Applied linguistics* 16(2). 141–158.
Carter, Ronald & Michael McCarthy. 2006a. *Cambridge grammar of English: a comprehensive guide: spoken and written English grammar and usage*. Cambridge: Cambridge University Press.

Carter, Ronald & Michael McCarthy. 2006b. *Cambridge Grammar of English: A Comprehensive Guide: Spoken and Written English Grammar and Usage*. Cambridge: Cambridge University Press.

Carter, Ronald & Michael McCarthy. 2017. Spoken Grammar: Where Are We and Where Are We Going? *Applied Linguistics* 38(1). 1–20. https://doi.org/10.1093/applin/amu080.

Catalán, ROSA Mª Jiménez & Rocío Mancebo Francisco. 2008. Vocabulary Input in EFL Textbooks. *Revista Española De Lingüística Aplicada (RESLA),(21)* 21. 147–166.

Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen (ed.), *Spoken and Written Language: Exploring Orality and Literacy* (Advances in Discourse Processes 9), 35–54. Norwood, NJ: Ablex.

Chambers, Angela. 2019. Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching* 52(4). 460–475. https://doi.org/10.1017/S0261444819000089.

Chapelle, Carol A. 2010. The spread of computer-assisted language learning. *Language Teaching* 43(1). 66–74. https://doi.org/10.1017/S0261444809005850.

Chen, Alvin Cheng-Hsien. 2016. A Critical Evaluation of Text Difficulty Development in ELT Textbook Series: A Corpus-Based Approach Using Variability Neighbor Clustering. *System* 58. 64–81. https://doi.org/10.1016/j.system.2016.03.011.

Chen, Alvin Cheng-Hsien. 2017. Assessing Text Difficulty Development in ELT Textbooks Series Using N-gram Language Models based on BNC. Presented at the Corpus Linguistics Conference 2017, Birmingham. https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper137.pdf (21 January, 2019).

Chen, Lin. 2010. An Investigation of Lexical Bundles in ESP Textbooks and Electrical Engineering Introductory Textbooks. In David Wood (ed.), *Perspectives on Formulaic Language: Acquisition and Communication*, 107–125. London: Continuum.

Cheng, Winnie. 2007. Sorry to Interrupt, But…: Pedagogical Implications of a Spoken Corpus. In Mari Carmen Campoy & María José Luzón (eds.), *Spoken Corpora in Applied Linguistics* (Linguistic Insights v. 51), 199–216. Bern: Peter Lang.

Cheng, Winnie, Christopher Greaves & Martin Warren. 2005. The creation of a prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic). *ICAME journal* 29. 47–68.

Cheng, Winnie & Martin Warren. 2005. // → Well I Have a DIFferent // ↘ THINking You Know //: A Corpus-Driven Study of Disagreement in Hong Kong Business Discourse. In M. Gotti & F. Bargiela (eds.), *Asian Business Discourse(s)*, 241–270. Frankfurt am Main: Peter Lang.

Cheng, Winnie & Martin Warren. 2006. I Would Say Be Very Careful Of…: Opine Markers in an Intercultural Business Corpus of Spoken English. In J. Bamford & M. Bondi (eds.), *Managing Interaction in Professional Discourse: Intercultural and Interdiscoursal Perspectives*, 46–57. Officina Edizioni. Rome.

Cheng, Winnie & Martin Warren. 2007. Checking Understandings: Comparing Textbooks and a Corpus of Spoken English in Hong Kong. *Language Awareness* 16(3). 190–207. https://doi.org/10.2167/la455.0.

Chicago Text Lab. 2020. US Novel Corpus. *Textual Optics Lab*. https://textual-optics-lab.uchicago.edu/us_novel_corpus (4 January, 2022).

Chien, Chu Ying & Kathie Young. 2007. The centrality of textbooks in teachers' work: Perceptions and use of textbooks in a Hong Kong primary school. *The Asia-Pacific Education Researcher*. De La Salle University 16(2). 155–163.

Chomsky, Noam. 1995. Language and Nature. *Mind*. [Oxford University Press, Mind Association] 104(413). 1–61.

Chomsky, Noam. 2002. *On Nature and Language*. (Ed.) Adriana Belletti & Luigi Rizzi. Cambridge: Cambridge University Press.

Chong, Chia Suan. 2012. The Teach-Off – My reaction to coursebooks & Uncount Nouns. *chiasuanchong.com*. https://chiasuanchong.com/2012/04/26/the-teach-off-my-reaction-to-coursebooks-uncountable-nouns/ (17 September, 2020).

Clark, Eve V. & Marisa Casillas. 2016. First language acquisition. In Keith Allan (ed.), *The Routledge Handbook of Linguistics*, 311–328. Abingdon; New York: Routledge.

Clarke, Isobelle & Dr Jack Grieve. 2017. Dimensions of Abusive Language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, 1–10. https://www.aclweb.org/anthology/W17-3001.pdf (11 February, 2021).

Cohen, Henri & Claire Lefebvre. 2017. *Handbook of Categorization in Cognitive Science*. 2nd edn. Amsterdam: Elsevier.

Collins, Peter. 2008. The progressive aspect in World Englishes: A corpus-based study. *Australian Journal of Linguistics* 28(2). 225–249.

Condon, Nora. 2008. How cognitive linguistic motivations influence the learning of phrasal verbs. In Frank Boers & Seth Lindstromberg (eds.), *Cognitive linguistic approaches to teaching vocabulary and phraseology*, 133–158. De Gruyter Mouton.

Conrad, Susan. 2000. Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century? *TESOL Quarterly* 34(3). 548–560. https://doi.org/10.2307/3587743.

Conrad, Susan. 2004. Corpus variety: Corpus linguistics, language variation, and language teaching. In John McH. Sinclair (ed.), *How to use corpora in language teaching* (Studies in Corpus Linguistics), vol. 12, 67–85. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.12.08con.

Conrad, Susan. 2013. Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In Susan Conrad & Douglas Biber (eds.), *Variation in English: Multi-Dimensional Studies*, 94–107. New York: Routledge.

Conrad, Susan & Douglas Biber (eds.). 2013. *Variation in English: Multi-Dimensional Studies* (Studies in Language and Linguistics). New York: Routledge.

Conrad, Susan, Douglas Biber & Geoffrey N. Leech. 2011. *Longman Student Grammar of Spoken and Written English. Workbook.* 11. impr. Harlow: Longman.

Conrad, Susan M. 1996a. *Academic discourse in two disciplines: Professional writing and student development in biology and history*. Northern Arizona University PhD dissertation.

Conrad, Susan M. 1996b. Investigating Academic Texts with Corpus-Based Techniques: An Example from Biology. *Linguistics and education* 8(3). 299–326.

Conseil supérieur des programmes. 2015. Projet de programme de cycle 4. https://www.education.gouv.fr/media/18161/download (1 January, 2018).

Consejería de Educación, Juventud y Deporte de Madrid. 2015. Decreto 48/2015 Curriculo de Educación Secundaria Obligatoria. Boletín oficial de la comunidad de Madrid. www.bocm.es (12 January, 2022).

Cook, Guy. 2000. *Language Play, Language Learning*. Oxford: Oxford University Press.

Cook, Vivian. 2002. The Functions of Invented Sentences: A Reply to Guy Cook. *Applied Linguistics* 23(2). 262–269.

Cools, Dorien & Lies Sercu. 2006. Die Beurteilung von Lehrwerken an Hand des Gemeinsamen Europäischen Referenzrahmens für Sprachen: Eine empirische Untersuchung von zwei kürzlich erschienenen Lehrwerken für Deutsch als Fremdsprache. *Zeitschrift für interkulturellen Fremdsprachenunterricht* 11(3). 1–20.

Council of Europe (ed.). 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR)*. Modern Languages Division, Strasbourg. Cambridge: Cambridge University Press.

Council of Europe (ed.). 2004. *Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR)*. 6. pr. Stuttgart: Klett.

Council of Europe (ed.). 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion volume with new descriptors*. Strasbourg: Council of Europe Publishing. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989 (6 February, 2022).

Council of Europe (ed.). 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion volume*. Strasbourg: Council of Europe Publishing. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4 (6 February, 2022).

Cowie, A. P. (ed.). 1998. *Phraseology: theory, analysis, and applications* (Oxford Studies in Lexicography and Lexicology). Oxford: Oxford University Press.

Crawford, William J. & Meixiu Zhang. 2021. How can register analysis inform task-based language teaching? *Register Studies*. John Benjamins 3(2). 180–206. https://doi.org/10.1075/rs.20021.cra.

Croft, William & D. Alan Cruse. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Crossley, Scott A., Kristopher Kyle & Ute Römer. 2019. Examining Lexical and Cohesion Differences in Discipline-Specific Writing Using Multi-Dimensional Analysis. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 189–216. Bloomsbury Academic.

Crossley, Scott, Laura K. Allen & Danielle McNamara. 2014. A Multi-Dimensional analysis of essay writing: What linguistic features tell us about situational parameters and the effects of language functions on judgments of quality. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Studies in Corpus Linguistics*, vol. 60, 197–238. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.60.07cro.

Crosthwaite, Peter (ed.). 2020. *Data-driven learning for the next generation: corpora and DDL for pre-tertiary learners*. London: Routledge.

Crowley, Tony. 2003. *Standard English and the politics of language*. 2nd ed. Houndmills, Basingstoke, Hampshire ; New York: Palgrave Macmillan.

Crystal, David. 2018. *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press.

Crystal, David & Derek Davy. 1969. *Investigating English Style* (English Language Series 1). 16. impr. Harlow: Longman.

Cullen, Richard & I-Chun (Vicky) Kuo. 2007. Spoken Grammar and ELT Course Materials: A Missing Link? *TESOL Quarterly* 41(2). 361–386. https://doi.org/10.1002/j.1545-7249.2007.tb00063.x.

Curry, Niall, Robbie Love & Olivia Goodman. in press. Adverbs on the move: Investigating publisher application of corpus research on recent language change to ELT coursebook development. *Corpora* 17(1). 1–38. https://doi.org/DOI: 10.3366/cor.2022.0233.

Dat, Bao. 2008. ELT Materials Used in Southeast Asia. In Brian Tomlinson (ed.), *English Language Learning Materials: A Critical Review*, 263–280. London: Continuum.

Dat, Bao. 2013. Developing Materials for Speaking Skills. In *Developing materials for language teaching*, 407–428. Second edition. London; New York: Bloomsbury.

Davies, M. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25(4). 447–464. https://doi.org/10.1093/llc/fqq018.

Davies, Mark. 2009. The 385+ million word *Corpus of Contemporary American English* (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14(2). 159–190. https://doi.org/10.1075/ijcl.14.2.02dav.

Davies, Mark. 2019. *The Movie Corpus*. Available online at https://www.english-corpora.org/movies/ (28 February, 2022).

Day, Richard R. & Julian Bamford. 1998. *Extensive Reading in the Second Language Classroom*. Cambridge: Cambridge University Press.

De Knop, Sabine & Gaëtanelle Gilquin. 2016. *Applied Construction Grammar*. Berlin; Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110458268.

Derwing, T. 2001. What speaking rates do non-native listeners prefer? *Applied Linguistics* 22(3). 324–337. https://doi.org/10.1093/applin/22.3.324.

Derwing, Tracey M. 1990. Speech Rate Is No Simple matter: Rate Adjustment and NS–NNS Communicative Success. *Studies in Second Language Acquisition* 12(3). 303–313. https://doi.org/10.1017/S0272263100009189.

Derwing, Tracey M., Ron I. Thomson, Jennifer A. Foote & Murray J. Munro. 2012. A Longitudinal Study of Listening Perception in Adult Learners of English: Implications for Teachers. *The Canadian Modern Language Review* 68(3). 247–266. https://doi.org/10.3138/cmlr.1215.

Desagulier, Guillaume. 2014. Visualizing distances in a set of near-synonyms: *Rather, quite, fairly* , and *pretty*. In Dylan Glynn & Justyna A. Robinson (eds.), *Human Cognitive Processing*, vol. 43, 145–178. Amsterdam: John Benjamins. https://doi.org/10.1075/hcp.43.06des.

Desagulier, Guillaume. 2017. *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics* (Quantitative Methods in the Humanities and Social Sciences). Cham: Springer International Publishing.

Deshors, Sandra C. 2017. Zooming in on verbs in the progressive: A collostructional and correspondence analysis approach. *Journal of English Linguistics* 45(3). 260–290.

Diehr, Bärbel (ed.). 2018. *Universitäre Englischlehrerbildung: Wege zu mehr Kohärenz im Studium und Korrespondenz mit der Praxis*. Berlin: Peter Lang. https://doi.org/10.3726/b14519.

Diehr, Bärbel. 2020. Kohärenz und Korrespondenz: Die fachdidaktische Perspektive auf die universitäre Englischlehrerbildung. In Michaela Heer & Ulrich Heinen (eds.), *Die Stimmen der Fächer hören: Fachprofil und Bildungsanspruch in der Lehrerbildung*, 325–342. Paderborn: Ferdinand Schöningh.

Diepenbroek, Lori G. & Tracey M. Derwing. 2014. To What Extent Do Popular ESL Textbooks Incorporate Oral Fluency and Pragmatic Development. *TESL Canada Journal* 30(7). 1. https://doi.org/10.18806/tesl.v30i7.1149.

Dinh, Thuy Ngoc & Fenty Lidya Siregar. 2021. Intercultural Competence and Parsnip: Voices From Teachers of English in Australia. In María Dolores López-Jiménez & Jorge Sánchez-Torres (eds.), *Intercultural Competence Past, Present and Future: Respecting the Past, Problems in the Present and Forging the Future* (Intercultural Communication and Language Education), 255–274. Singapore: Springer.

Dirven, René. 1990. Pedagogical grammar. *Language Teaching* 23(01). 1. https://doi.org/10.1017/S0261444800005498.

Dirven, René & Günter Radden. 1977. *Semantische Syntax des Englischen*. Vol. 13. Wies: Athenaion.

Diwersy, Sascha, Stephanie Evert & Stella Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*, 174–204. Berlin, Boston: De Gruyter.

Diz-Otero, Mario, Iago Portela-Pino, Sara Domínguez-Lloria & Margarita Pino-Juste. 2022. Digital competence in secondary education teachers during the COVID-19-derived pandemic: comparative analysis. *Education + Training* ahead-of-print(ahead-of-print). https://doi.org/10.1108/ET-01-2022-0001.

Dörnyei, Zoltán, Valerie Durow & Khawla Zahran. 2004. Individual differences and their effects on formulaic sequence acquisition. In Norbert Schmitt (ed.), *Formulaic sequences: acquisition, processing, and use* (Language Learning and Language Teaching v. 9), 87–106. Amsterdam; Philadelphia: John Benjamins.

Duffley, Patrick J. 2006. *The English Gerund-Participle: A Comparison with the Infinitive* (Berkeley Insights in Linguistics and Semiotics). Vol. 61. Frankfurt: Peter Lang.

Dumont, Amandine. 2018. *Fluency and disfluency: A corpus study of non-native and native speaker (dis)fluency profiles*. Louvain: Université catholique de Louvain PhD dissertation. http://hdl.handle.net/2078.1/198393.

Eckhardt, Suzanne. 2001. *Reported speech : empirical corpus findings compared with EFL/ESL textbook presentations*. Ames: Iowa State University Unpublished M.A. thesis. https://doi.org/10.31274/rtd-180813-8333. https://lib.dr.iastate.edu/rtd/17551/ (30 January, 2020).

Edwards, Alison. 2014. *English in the Netherlands: Functions, forms and attitudes*. University of Cambridge PhD dissertation.

Edwards, Alison. 2016. *English in the Netherlands: Functions, forms and attitudes* (Varieties of English Around the World). Vol. G56. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/veaw.g56. http://www.jbe-platform.com/content/books/9789027267207 (28 January, 2022).

Egbert, Jesse & Douglas Biber. 2018. Do all roads lead to Rome?: Modeling register variation with factor analysis and discriminant analysis. *Corpus Linguistics and Linguistic Theory* 14(2). 233–273. https://doi.org/10.1515/cllt-2016-0016.

Egbert, Jesse, Douglas Biber & Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology* 66(9). 1817–1831. https://doi.org/10.1002/asi.23308.

Egbert, Jesse, Tove Larsson & Douglas Biber. 2020. *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. 1st edn. Cambridge University Press. https://doi.org/10.1017/9781108888790. https://www.cambridge.org/core/product/identifier/9781108888790/type/element (5 November, 2020).

Egbert, Jesse & Michaela Mahlberg. 2020. Fiction – one register or two?: Speech and narration in novels. *Register Studies* 2(1). 72–101. https://doi.org/10.1075/rs.19006.egb.

Egbert, Jesse & Shelley Staples. 2019. Doing Multi-Dimensional Analysis in SPSS, SAS, and R. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 125–144. Bloomsbury Academic. https://doi.org/10.5040/9781350023857.

Elio, Renee & John R Anderson. 1981. The Effects of Category Generalizations and Instance Similarity on Schema Abstraction. *ournal of Experimental Psychology: Human Learning & Memory* 7(6). 397–417.

Elio, Renee & John R. Anderson. 1984. The effects of information order and learning mode on schema abstraction. *Memory & Cognition* 12(1). 20–30. https://doi.org/10.3758/BF03196994.

Ellis, N. C. & D. Larsen-Freeman. 2006. Language Emergence: Implications for Applied Linguistics--Introduction to the Special Issue. *Applied Linguistics* 27(4). 558–589. https://doi.org/10.1093/applin/aml028.

Ellis, Nick C. 1998. Emergentism, Connectionism and Language Learning. *Language Learning* 48(4). 631–664. https://doi.org/10.1111/0023-8333.00063.

Ellis, Nick C. 2002. Frequency effects in language processing. *Studies in second language acquisition* 24(2). 143–188.

Ellis, Nick C. 2006. Language Acquisition as Rational Contingency Learning. *Applied Linguistics* 27(1). 1–24. https://doi.org/10.1093/applin/ami038.

Ellis, Nick C. 2008. Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 endstate. In Peter Robinson & Nick C. Ellis (eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*, 372–405. London: Routledge.

Ellis, Nick C. 2019. Essentials of a Theory of Language Cognition. *The Modern Language Journal* 103(S1). 39–60. https://doi.org/10.1111/modl.12532.

Ellis, Nick C. & Fernando Ferreira-Junior. 2009a. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7(1). 188–221.

Ellis, Nick C. & Fernando Ferreira-Junior. 2009b. Construction Learning as a Function of Frequency, Frequency Distribution, and Function. *The Modern Language Journal* 93(3). 370–385. https://doi.org/10.1111/j.1540-4781.2009.00896.x.

Ellis, Nick C. & Diane Larsen-Freeman. 2009. Constructing a Second Language: Analyses and Computational Simulations of the Emergence of Linguistic Constructions From Usage. *Language Learning* 59. 90–125. https://doi.org/10.1111/j.1467-9922.2009.00537.x.

Ellis, Nick C., Ute Römer & Matthew Brook O'Donnell. 2016. *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar* (Language Learning Monograph Series). Malden, MA: Wiley-Blackwell.

Ellis, Rod. 1989. Are classroom and naturalistic acquisition the same?: A study of the classroom acquisition of German word order rules. *Studies in second language Acquisition* 11(3). 305–328.

Ellis, Rod. 1997. The empirical evaluation of language teaching materials. *ELT journal* 51(1). 36–42.

Elman, Jeffrey L, Elizabeth A Bates, Mark H Johnson, Annette Karmiloff-Smith, Domenico Parisi & Kim Plunkett. 1996. *Rethinking innateness: A connectionist perspective on development*. Vol. 10. Cambridge, MA: MIT press.

European Commission, EACEA, & Eurydice. 2017. *Key data on teaching languages at school in Europe: Eurydice Report*. Luxemburg: Publications Office of the European Union. http://data.europa.eu/doi/10.2797/839825 (15 October, 2018).

Evert, Stephanie. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Institut für maschinelle Sprachverarbeitung, University of Stuttgart PhD thesis.

Evert, Stephanie. 2009. Corpora and collocations. In Anke Lüdeling & Merja Kytö (eds.), *Handbooks of Linguistics and Communication Science*. Berlin, New York: Mouton de Gruyter. https://doi.org/10.1515/9783110213881.2.1212.

Evert, Stephanie. 2018. Statistics for Linguists with R – A SIGIL Course: Unit 7: A multivariate approach to linguistic variation. FAU Erlangen-Nürnberg. http://www.stephanie-evert.de/SIGIL/sigil_R/ (4 November, 2021).

Fankhauser, Anna. forthcoming. Suggestions for a new model of functional phraseme categorization for applied purposes. In Johanna Monti, Ruslan Mitkov & Gloria Corpas Pastor (eds.), *Recent Advances in Multiword Units in Machine Translation and Translation Technology*. Amsterdam: John Benjamins.

Fankhauser, Anna. in preparation. *Formulaic Language in the EFL Classroom: A Corpus-Based Study of Phraseological Items in British English and American English Conversation with Implications for EFL Teaching*. Osnabrück University.

Feierabend, Sabine, Thomas Rathgeb, Hediye Kheredmand & Stephan Glöckler. 2020. *JIM-Studie 2020: Jugend, Information, Medien. Basisuntersuchung zum Medienumgang 12- bis 19-Jähriger*. Medienpädagogischer Forschungs- verbund Südwest (mpfs). https://www.mpfs.de/fileadmin/files/Studien/JIM/2020/JIM-Studie-2020_Web_final.pdf (20 January, 2022).

Fogal, Gary G. & Marjolijn H. Verspoor. 2020. *Complex Dynamic Systems Theory and L2 Writing Development*. Amsterdam: John Benjamins.

Fominykh, Mikhail, Elizaveta Shikhova, Maria Victoria Soule, Maria Perifanou & Daria Zhukova. 2021. Digital Competence Assessment Survey for Language Teachers. In Panayiotis Zaphiris & Andri Ioannou (eds.), *Learning and Collaboration Technologies: New Challenges and Learning Experiences*, vol. 12784, 264–282. Cham: Springer. (17 February, 2022).

Forsberg, Julia, Susanne Mohr & Sandra Jansen. 2019. "The goal is to enable students to communicate": Communicative competence and target varieties in TEFL practices in Sweden and Germany. *European Journal of Applied Linguistics* 7(1). 31–60. https://doi.org/10.1515/eujal-2018-0006.

Fournier, Yann, Anne Gaudry-Lachet & DEPP-MIREI. 2017. *L'apprentissage des langues vivantes étrangères dans l'Union européenne : parcours des élèves*. Note d'information. Ministère de l'Education Nationale. https://www.education.gouv.fr/l-apprentissage-des-langues-vivantes-etrangeres-dans-l-union-europeenne-parcours-des-eleves-2588 (1 August, 2018).

Frazier, Stefan. 2003. A Corpus Analysis of Would-Clauses without Adjacent If-Clauses. *TESOL Quarterly* 37(3). 443–466. https://doi.org/10.2307/3588399.

Freeman, David & Susan Holden. 1986. Authentic listening materials. *Techniques of Teaching. London: Modern English Publications* 67–69.

Freudenstein, Reinhold. 2002. Was morgen geschah... Schulischer Fremdsprachenunterricht gestern, heute - und in Zukunft? In Christiane Neveling (ed.), *Perspektiven für die zukünftige Fremdsprachendidaktik*, 45–62. Tübingen: Gunter Narr Verlag.

Friederici, Luisa. 2019. Vorschlag für eine pluriperspektivische Analyse der Zielgruppe zur Auswahl eines neuen Lehrwerks oder: Vergesst die Lehrer nicht! *Pandaemonium Germanicum* 22. 281–301. https://doi.org/10.11606/1982-88372237281.

Friginal, Eric. 2018. *Corpus linguistics for English teachers: new tools, online resources, and classroom activities*. New York, NY: Routledge. https://doi.org/10.4324/9781315649054.

Friginal, Eric & Jack A. Hardy. 2014. Conducting Multi-Dimensional Analysis Using SPSS. In Tony Berber Sardinha & Douglas Biber (eds.), *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber* (Studies in Corpus Linguistics (SCL) 60), 297–316. Amsterdam: John Benjamins.

Friginal, Eric & Sara Weigle. 2014. Exploring multiple profiles of L2 writing using multi-dimensional analysis. *Journal of Second Language Writing* 26. 80–95. https://doi.org/10.1016/j.jslw.2014.09.007.

Fuchs, Robert & Ulrike Gut. 2015. An apparent time study of the progressive in Nigerian English. In Peter Collins (ed.), *Grammatical Change in English World-Wide*, 373–387. Amsterdam: John Benjamins.

Fuchs, Robert & Valentin Werner. 2018. The use of stative progressives by school-age learners of English and the importance of the variable context: Myth vs. (corpus) reality. *International Journal of Learner Corpus Research* 4(2). 195–224. https://doi.org/10.1075/ijlcr.17010.fuc.

Fujimoto, Kazuko. 2017. Do English Textbooks Reflect the Actual Use of English?: The Present Perfect and Temporal Adverbials. Presented at the The 9th International Corpus Linguistics Conference, University of Birmingham. https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper117.pdf (9 January, 2017).

Gabrielatos, Costas. 1994. *Collocations: Pedagogical Implications, and Their Treatment in Pedagogical Materials*. Research Centre for English and Applied Linguistics, Cambridge University, UK. Unpublished PhD thesis. https://www.researchgate.net/publication/261708736_Collocations_Pedagogical_implications_and_their_treatment_in_pedagogical_materials.

Gabrielatos, Costas. 2003. Conditional Sentences: ELT Typology and Corpus Evidence. Presented at the BAAL 36th Annual Meeting. https://www.researchgate.net/publication/261708834_Conditional_Sentences_ELT_typology_and_corpus_evidence.

Gabrielatos, Costas. 2006. Corpus-Based Analysis of Pedagogical Materials: If-Conditionals in ELT Coursebooks and the BNC. Presented at the 7th Teaching and Language Corpora Conference. https://www.researchgate.net/publication/228880683_Corpus-based_evaluation_of_pedagogical_materials_If-conditionals_in_ELT_coursebooks_and_the_BNC (6 February, 2020).

Gabrielatos, Costas. 2013. *If*-conditionals in ICLE and the BNC: A success story for teaching or learning? In S. Granger, F. Meunier & G. Gilquin (eds.), *Twenty Years of Learner Corpus Research: Looking back, moving ahead*, 155–166. Louvain-la-Neuve: Presses Universitaires de Louvain.

Gabrielatos, Costas. 2019. *If* -Conditionals and Modality: Frequency Patterns and Theoretical Explanations. *Journal of English Linguistics* 47(4). 301–334. https://doi.org/10.1177/0075424219875994.

Gardner, Dee & Mark Davies. 2007. Pointing Out Frequent Phrasal Verbs: A Corpus-Based Analysis. *TESOL Quarterly* 41(2). 339–359. https://doi.org/10.1002/j.1545-7249.2007.tb00062.x.

Gardner, Sheena, Hilary Nesi & Douglas Biber. 2019. Discipline, Level, Genre: Integrating Situational Perspectives in a New MD Analysis of University Student Writing. *Applied Linguistics* 40(4). 646–674. https://doi.org/10.1093/applin/amy005.

Garinger, Dawn. 2002. Textbook selection for the ESL classroom. *Center for Applied Linguistics Digest*. Retrieved from http://www.mcael.org/uploads/File/provider_library/Textbook_Eval_CAL.pdf (12 January, 2017).

Garton, Sue & Kathleen Graves (eds.). 2014. *International Perspectives on Materials in ELT*. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137023315. http://link.springer.com/10.1057/9781137023315 (8 April, 2020).

Gass, Susan M. 1997. *Input, Interaction, and the Second Language Learner*. New York: Routledge. (11 January, 2022).

Geeraerts, Dirk. 2006. *Cognitive Linguistics: Basic Readings*. Berlin: Walter de Gruyter.

Gehring, Wolfgang. 2013. Can't judge a book by its cover: An analytical approach to textbook innovations. In Maria Eisenmann & Theresa Summer (eds.), *Basic Issues in EFL Teaching and Learning*, 357–370. Second Edition. Heidelberg: Winter.

Gezelter, Dan. 2009. Being Scientific: Falsifiability, Verifiability, Empirical Tests, and Reproducibility | The OpenScience Project. *The OpenScience Project*. https://openscience.org/being-scientific-fasifiability-verifiability-empirical-tests-and-reproducibility/ (21 August, 2021).

Ghanbari, Nasim, Fatemeh Esmaili & Mohammad Reza Shamsaddini. 2015. The Effect of Using Authentic Materials on Iranian EFL Learners' Vocabulary Learning. *Theory and Practice in Language Studies* 5(12). 2459–2468. https://doi.org/10.17507/tpls.0512.05.

Ghosn, Irma K. 2013. Talking like Texts and Talking about Texts: How Some Primary School Coursebook Tasks are Realized in the Classroom. In *Developing materials for language teaching*, 291–305. Second edition. London; New York: Bloomsbury.

Gießing, Jürgen. 2004. Zankapfel "Lehrbuch": für und wider ein etabliertes Unterrichtsmittel. *PRAXIS Fremdsprachenunterricht* (2). 82–84.

Gilmore, Alex. 2004. A Comparison of Textbook and Authentic Interactions. *ELT Journal* 58(4). 363–374. https://doi.org/10.1093/elt/58.4.363.

Gilmore, Alex. 2007. Authentic materials and authenticity in foreign language learning. *Language Teaching* 40(02). 97. https://doi.org/10.1017/S0261444807004144.

Gilmore, Alex. 2011. "I Prefer Not Text": Developing Japanese Learners' Communicative Competence with Authentic Materials. *Language Learning* 61(3). 786–819. https://doi.org/10.1111/j.1467-9922.2011.00634.x.

Gilmore, Alex. 2019. Materials and authenticity in language teaching. In Steve Walsh & Steve Mann (eds.), *The Routledge Handbook of English Language Teacher Education*, 299–318. 1st edn. London; New York, NY: Routledge. https://doi.org/10.4324/9781315659824-21.

Gilquin, Gaëtanelle. 2012. Lexical infelicity in causative constructions. Comparing native and learner collostructions. In Jaakko Leino & Ruprecht von Waldenfels (eds.), *Analytical causatives: from 'give' and 'come' to 'let' and 'make,'* 41–64. München: Lincom Europa.

Gilquin, Gaëtanelle. 2015a. Contrastive Collostructional Analysis: Causative Constructions in English and French. *Zeitschrift für Anglistik und Amerikanistik* 63(3). https://doi.org/10.1515/zaa-2015-0022.

Gilquin, Gaëtanelle. 2015b. The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory* 11(1). https://doi.org/10.1515/cllt-2014-0005.

Gilquin, Gaëtanelle. 2016a. Input-dependent L2 acquisition: Causative constructions in English as a foreign and second language. In Sabine De Knop & Gaëtanelle Gilquin (eds.), *Applied Construction Grammar*, 198–259. Berlin, Boston: De Gruyter.

Gilquin, Gaëtanelle. 2016b. Discourse markers in L2 English: From classroom to naturalistic input. In Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja & Sarah Chevalier (eds.), *Studies in Language Companion Series*, vol. 177, 213–249. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.177.09gil.

Gilquin, Gaëtanelle. 2018. American and/or British influence on L2 Englishes – Does context tip the scale(s)? In Sandra C. Deshors (ed.), *Varieties of English Around the World*, vol. G61, 187–216. Amsterdam: John Benjamins.

Gilquin, Gaëtanelle & Samantha Laporte. 2021. The use of online writing tools by learners of English: Evidence from a process corpus. *International Journal of Lexicography* 34(4). 472–492. https://doi.org/10.1093/ijl/ecab012.

Gnutzmann, Claus (ed.). 1999. *Teaching and learning English as a global language: native and non-native perspectives* (ZAA Studies 8). Tübingen: Stauffenburg-Verl.

Gnutzmann, Claus & Frauke Intemann (eds.). 2008. *The Globalisation of English and the English Language Classroom*. Tübingen: Narr Francke.

Goldberg, Adele. 1996. Making one's way through the data. In Masayoshi Shibatani & Sandra A. Thompson (eds.), *Grammatical constructions: their form and meaning*, 29–53. 2nd edn. 1999. Oxford [England] : New York: Clarendon Press ; Oxford University Press.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press.

Goldberg, Adele E., Devin M. Casenhiser & Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive Linguistics* 15(3). https://doi.org/10.1515/cogl.2004.011.

Gorsuch, Richard L. 2014. *Factor Analysis*. 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates. https://doi.org/10.4324/9781315735740.

Götz, Sandra. 2013. *Fluency in Native And Nonnative English Speech* (Studies in Corpus Linguistics volume 53). Amsterdam: John Benjamins.

Goulart, Larissa, Bethany Gray, Shelley Staples, Amanda Black, Aisha Shelton, Douglas Biber, Jesse Egbert & Stacey Wizner. 2020. Linguistic Perspectives on Register. *Annual Review of Linguistics* 6(1). 435–455. https://doi.org/10.1146/annurev-linguistics-011718-012644.

Goulart, Larissa & Margaret Wood. 2021. Methodological synthesis of research using multi-dimensional analysis. *Journal of Research Design and Statistics in Linguistics and Communication Science* 6(2). 107–137. https://doi.org/10.1558/jrds.18454.

Gouverneur, Céline. 2008a. The Phraseological Patterns of High-frequency Verbs in Advanced English for General Purposes: A Corpus-driven Approach to EFL Textbook Analysis. In Fanny Meunier & Sylviane Granger (eds.), *Phraseology in Foreign Language Learning and Teaching*, 223–243. Amsterdam: John Benjamins.

Gouverneur, Céline. 2008b. The phraseological patterns of high-frequency verbs in advanced English for general purposes: A corpus-driven approach to EFL textbook analysis. In Fanny Meunier & Sylviane Granger (eds.), *Phraseology in foreign language learning and teaching*, 223–243.

Grabowski, Łukasz. 2015. Keywords and lexical bundles within English pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes* 38. 23–33. https://doi.org/10.1016/j.esp.2014.10.004.

Granger, Sylviane. 1998. Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In A. P. Cowie (ed.), *Phraseology: Theory, Analysis, and Applications*. Oxford; New York: Clarendon Press; Oxford University Press.

Granger, Sylviane. 2004. Computer learner corpus research: current status and future prospects. *Language and Computers* 52(1). 123–145.

Granger, Sylviane. 2015. The contribution of learner corpora to reference and instructional materials design. In Sylviane Granger, Gaetanelle Gilquin & Fanny Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, 485–510. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.022.

Granger, Sylviane. 2018. Has Lexicography Reaped the Full Benefit of the (Learner) Corpus Revolution? In Simon Krek, Jaka Čibej, Vojko Gorjanc & Iztok Kosem (eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, 17–24. Ljubljana: Faculty of Arts, University of Ljubljana. https://doi.org/10.4312/9789610600961.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier & Magali Paquot. 2009. *International Corpus of Learner English*. Louvain-la-Neuve: UCL Presses Universitaires de Louvain.

Granger, Sylviane, Joseph Hung & Stephanie Petch-Tyson (eds.). 2002. *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*. Amsterdam: John Benjamins.

Grau, Maike. 2009. Worlds apart? English in German youth cultures and in educational settings. *World Englishes* 28(2). 160–174. https://doi.org/10.1111/j.1467-971X.2009.01581.x.

Graves, Kathleen & Sue Garton. 2019. Materials use and development. In Steve Walsh & Steve Mann (eds.), *The Routledge Handbook of English Language Teacher Education*, 417–431. 1st edn. London; New York, NY: Routledge.

Gray, Bethany. 2015. *Linguistic Variation in Research Articles: when discipline tells only part of the story* (Studies in Corpus Linguistics 71). Amsterdam: John Benjamins.

Gray, Bethany. 2019. Tagging and counting linguistic features for multi-dimensional analysis. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 43–66. London: Bloomsbury Academic. https://doi.org/10.5040/9781350023857.

Gray, Bethany & Jesse Egbert. 2019. Editorial: Register and register variation. *Register Studies* 1(1). 1–9. https://doi.org/10.1075/rs.00001.edi.

Gray, Bethany & Jesse Egbert. 2021. Register in L1 and L2 language development. *Register Studies* 3(2). 177–179. https://doi.org/10.1075/rs.21010.gra.

Gray, John. 2000. The ELT coursebook as cultural artefact: how teachers censor and adapt. *ELT Journal* 54(3). 274–283. https://doi.org/10.1093/elt/54.3.274.

Gray, John. 2002. The global coursebook in English language teaching. In D. Block & D. Cameron (eds.), *Globalization and language teaching*, 151–167. Routledge.

Gray, John. 2010. *The Construction of English: Culture, Consumerism and Promotion in the Elt Global Coursebook*. Basingstoke: Palgrave Macmillan.

Gries, Stefan Th. 2005. Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1(2). https://doi.org/10.1515/cllt.2005.1.2.277.

Gries, Stefan Th. 2012. Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language* 36(3). 477–510.

Gries, Stefan Th. 2013. Statistical tests for the analysis of learner corpus data. In Ana Díaz-Negrillo, Nicolas Ballier & Paul Thompson (eds.), *Studies in Corpus Linguistics*, vol. 59, 287–310. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.59.17gri.

Gries, Stefan Th. 2015a. More (old and new) misunderstandings of collostructional analysis: On Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3). https://doi.org/10.1515/cog-2014-0092.

Gries, Stefan Th. 2015b. The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora* 10(1). 95–125. https://doi.org/10.3366/cor.2015.0068.

Gries, Stefan Th. 2018. On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies* 1(2). 276–308. https://doi.org/10.1075/jsls.00005.gri.

Gries, Stefan Th. 2019a. 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics* 24(3). 385–412. https://doi.org/10.1075/ijcl.00011.gri.

Gries, Stefan Th. 2019b. *Coll.analysis 3.5 package for R*. http://www.stgries.info/teaching/groningen/ (14 February, 2019).

Gries, Stefan Th. & Sandra C. Deshors. 2020. Statistical analyses of learner corpus data. In Nancy Tracy-Ventura & Magali Paquot (eds.), *Routledge Handbook of SLA and Corpora*, 119–132. New York: Routledge.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4). 635–676.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2010. Converging evidence II: More on the association of verbs and constructions. In John Newman & Sally Rice (eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*, 59–72. CSLI Publications, Center for the Study of Language and Information.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004a. Extending collostructional analysis: A corpus-based perspective on "alternations." *International Journal of Corpus Linguistics* 9(1). 97–129. https://doi.org/10.1075/ijcl.9.1.06gri.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004b. Covarying Collexemes in the Into-causative. *International Journal of Corpus Linguistics* 1(9). 97–129.

Grieve, Jack, Douglas Biber, Eric Friginal & Tatiana Nekrasova. 2010. Variation Among Blogs: A Multi-dimensional Analysis. In Alexander Mehler, Serge Sharoff & Marina Santini (eds.), *Genres on the Web*, vol. 42, 303–322. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-9178-9_14.

Grieve-Smith, Angus B. 2007. The envelope of variation in multidimensional register and genre analyses. In Eileen Fitzpatrick (ed.), *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*, 21–42. Amsterdam: Brill Rodopi.

Griffiths, Roger. 1990. Speech Rate and NNS Comprehension: A Preliminary Study in Time-Benefit Analysis. *Language Learning* 40(3). 311–336. https://doi.org/10.1111/j.1467-1770.1990.tb00666.x.

Griffiths, Roger. 1991. Language Classroom Speech Rates: A Descriptive Study. *TESOL Quarterly* 25(1). 189. https://doi.org/10.2307/3587050.

Gurzynski-Weiss, Laura, Kimberly L Geeslin, Danielle Daidone, Bret Linford, Avizia Yim Long, Ian Michalski & Megan Solon. 2018. Examining multifaceted sources of input. Variationist and usage-based approaches to understanding the L2 classroom. In *Usage-Inspired L2 Instruction: Researched Pedagogy*, vol. 49, 291–311. Amsterdam; Philadelphia: John Benjamins.

Gut, Ulrike & Robert Fuchs. 2013. Progressive Aspect in Nigerian English. *Journal of English Linguistics* 41(3). 243–267. https://doi.org/10.1177/0075424213492799.

Hahn, Angela, Sabine Reich & Josef Schmied. 2000. Aspect in the Chemnitz Internet Grammar. In Christian Mair & Marianne Hundt (eds.), *Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999*, 131–139. Amsterdam: Rodopi.

Hair, Joseph F., William C. Black, Barry J. Babin & Rolph E. Anderson. 2019. *Multivariate data analysis*. Eighth edition. Andover, Hampshire: Cengage.

Hall Christopher J, Wicaksono Rachel, Liu Shu, Qian Yuan & Xu Xiaoqing. 2017. Exploring teachers' ontologies of English: Monolithic conceptions of grammar in a group of Chinese teachers. *International Journal of Applied Linguistics* 27(1). 87–109. https://doi.org/10.1111/ijal.12107.

Hallet, Wolfgang & Michael K Legutke. 2013. Task-approaches Revisited: New Orientations, New Perspectives. *The European Journal of Applied Linguistics and TEFL* 2(2). 139–159.

Halliday, Michael A. K. 1993. *Language as social semiotic: The social interpretation of language and meaning*. 1. publ. in paperback, 8. impr. London: Arnold.

Halliday, Michael AK. 1988. On the language of physical science. In M. Ghadessy (ed.), *Registers of written English: Situational factors and linguistic features*, 162–172.

Hardie, Andrew. 2014. Modest XML for Corpora: Not a standard, but a suggestion. *ICAME Journal* 38(1). 73–103.

Harel, Idit & Seymour Papert. 1991. *Constructionism: Research Reports and Essays, 1985-1990*. Ablex Publishing Corporation.

Harwood, Nigel. 2005. What do we want EAP teaching materials for? *Journal of English for Academic Purposes* 4(2). 149–161.

Hausmann, Franz Josef. 2005. Isotopie, scénario, collocation et exemple lexicographique. In Michaela Heinz (ed.), *L'exemple lexicographique dans les dictionnaires français contemporains*, 283–292. De Gruyter. https://doi.org/10.1515/9783110924466.283.

Henry, Alastair. 2014. Swedish students' beliefs about learning English in and outside of school. In David Lasagabaster, Aintzane Doiz & Juan Manuel Sierra (eds.), *Language Learning & Language Teaching*, vol. 40, 93–116. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.40.05hen.

Henson, Robin K. & J. Kyle Roberts. 2006. Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice. *Educational and Psychological Measurement* 66(3). 393–416. https://doi.org/10.1177/0013164405282485.

Herbst, Thomas. 1996. What are collocations: Sandy beaches or false teeth? *English Studies* 77(4). 379–393. https://doi.org/10.1080/00138389608599038.

Herbst, Thomas. 2016. Foreign language learning is construction learning–what else? Moving towards Pedagogical Construction Grammar. In Sabine De Knop & Gaëtanelle Gilquin (eds.), *Applied Construction Grammar*, vol. 32, 56–96. Berlin; Boston: De Gruyter Mouton.

Herbst, Thomas, David Heath & Ian F Roe. 2013. *Valency Dictionary of English: a Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives.* Tubingen: De Gruyter Mouton. https://doi.org/10.1515/9783110892581.

Herbst, Thomas, Hans-Jörg Schmid & Susen Faulhaber (eds.). 2014. *Constructions, collocations, patterns* (Trends in Linguistics. Studies and Monographs 282). Berlin; Boston: Walter De Gruyter.

Hermes, Liesel. 2009. "Reading can be fun if…": Lektüren in der Sekundarstufe I. In Jan Hollm (ed.), *Literaturdidaktik und Literaturvermittlung im Englischunterricht der Sekundarstufe I*, 7–22. Trier: Wissenschaftlicher Verlag Trier.

Hessisches Kultusministerium. 2010. Lehrplan Englisch - Gymnasialer Bildungsgang - Jahrgangsstufen 5G bis 9G. https://kultusministerium.hessen.de/sites/kultusministerium.hessen.de/files/2021-06/g8-englisch.pdf (6 February, 2022).

Hoey, Michael. 2005. *Lexical Priming: A New Theory of Words and Language*. Routledge.

Holliday, Adrian. 2005. *The Struggle to teach English as an international language*. Oxford; New York: Oxford University Press.

Holmes, Janet. 1988. Doubt and Certainty in ESL Textbooks. *Applied Linguistics* 9(1). 21–44. https://doi.org/10.1093/applin/9.1.21.

Honnibal, Matthew & Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1373–1378. Lisbon, Portugal: Association for Computational Linguistics.

Honnibal, Matthew & Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear* 7.

Hu, Chieh–Fang & Cheyenne Maechtle. 2021. Construction Learning by Child Learners of Foreign Language: Input Distribution and Learner Factors. *The Modern Language Journal* 105(1). 335–354. https://doi.org/10.1111/modl.12698.

Huang, Pingping. 2019. Textbook interaction: A study of the language and cultural contextualisation of English learning textbooks. *Learning, Culture and Social Interaction* 21. 87–99. https://doi.org/10.1016/j.lcsi.2019.02.006.

Hughes, Rebecca. 2010. What a corpus tells us about grammar teaching materials. In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*, 401–412. Abingdon, Oxon; New York: Routledge. https://doi.org/10.4324/9780203856949.ch29.

Hundt, Marianne. 2004. The passival and the progressive passive: A case study of layering in the English aspect and voice systems. In Hans Lindquist & Christian Mair (eds.), *Corpus Approaches to Grammaticalization in English* (Studies in Corpus Linguistics 13), 79–120. Amsterdam: John Benjamins.

Hundt, Marianne, Andrea Sand & Rainer Siemund. 1998. Manual of information to accompany the Freiburg-LOB Corpus of British English ('FLOB'). Albert-Ludwigs-Universität Freiburg. http://korpus.uib.no/icame/frown/ (23 January, 2022).

Hundt, Marianne, Andrea Sand & Paul Skandera. 1999. Manual of information to accompany the Freiburg-Brown Corpus of American English. Albert-Ludwigs-Universität Freiburg. http://korpus.uib.no/icame/flob/ (23 January, 2022).

Hundt, Marianne & Katrin Vogel. 2011. Overuse of the progressive in ESL and learner Englishes – fact or fiction? In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*, 145–165. Amsterdam: John Benjamins.

Hunston, Susan. 2002. *Corpora in Applied Linguistics* (The Cambridge applied linguistics series). Cambridge: Cambridge University Press.

Hunston, Susan & Gill Francis. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English* (Studies in Corpus Linguistics 4). Amsterdam; Philadelphia: John Benjamins.

Husson, François, Pierre-André Cornillon, Arnaud Guyader, Nicolas Jégou, Julie Josse, Nicolas Klutchnikoff, Erwan Le Pennec, Eric Matzner-Løber, Laurent Rouvière & Benoît Thieurmel. 2018. *R pour la statistique et la science des données*. Rennes: Presses universitaires de Rennes.

Husson, François, Sébastien Lê & Jérôme Pagès. 2017. *Exploratory Multivariate Analysis by Example Using R*. 2nd edn. Boca Raton: Chapman and Hall/CRC Press. https://doi.org/10.1201/b21874.

Hyland, Ken. 1994. Hedging in academic writing and EAP textbooks. *English for Specific Purposes* 13(3). 239–256.

Hyland, Ken. 2013. Materials for Developing Writing Skills. In Brian Tomlinson (ed.), *Developing Materials for Language Teaching*, 391–406. Second Edition. London; New York: Bloomsbury.

Hymes, Dell. 1984. Sociolinguistics: Stability and consolidation. *International Journal of the Sociology of Language*. Mouton Publishers 45. 39–45.

Ide, Nancy. 1996. *Corpus Encoding Standard*. Expert Advisory Group on Language Engineering Standards (EAGLES). https://www.cs.vassar.edu/CES/ (2 January, 2022).

Israel, Michael. 1996. The Way Constructions Grow. In Adele Goldberg (ed.), *Conceptual structure, discourse and language*, 217–230. Stanford: CSLI.

Jablonkai, Reka R. & Eniko Csomay. 2022. *The Routledge Handbook of Corpora and English Language Teaching and Learning*. London: Routledge.

Jacobs, G. M. & J. Ball. 1996. An investigation of the structure of group activities in ELT coursebooks. *ELT Journal* 50(2). 99–107. https://doi.org/10.1093/elt/50.2.99.

Jansen, Sandra, Susanne Mohr & Julia Forsberg. 2021. Standard language ideology in the English language classroom: Suggestions for EIL-informed teacher education. In Marcus Callies, Stefanie Hehner, Philipp Meer & Micahel Westphal (eds.), *Glocalising Teaching English as an International Language*. Routledge.

Jenkins, Jennifer. 1998. Which pronunciation norms and models for English as an International Language? *ELT Journal* 52(2). 119–126. https://doi.org/10.1093/elt/52.2.119.

Jenkins, Jennifer. 2000. *The phonology of English as an international language*. Oxford: Oxford University Press.

Jenkins, Jennifer. 2003. *World Englishes: a resource book for students* (Routledge English Language Introductions). New York: Routledge.

Jiang, Xiangying. 2006. Suggestions: What Should ESL Students Know? *System* 34(1). 36–54. https://doi.org/10.1016/j.system.2005.02.003.

Johansson, Stig, Geoffrey N Leech & Helen Goodluck. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computer*. Department of English, University of Oslo.

Johns, Tim. 1986. Micro-concord: A language learner's research tool. *System*. Elsevier 14(2). 151–162.

Johns, Tim. 1993. Data-driven learning: An update. *TELL & CALL* 3. 23–32.

Johns, Tim. 2002. Data-driven learning: The perpetual challenge. In Bernhard Kettemann & Georg Marko (eds.), *Teaching and Learning by Doing Corpus Analysis*, 105–117. Leiden: Brill.

Johns, Tim. 2014. Contexts: The background, development and trialling of a concordance-based CALL program. In Anne Wichmann & Steven Fligelstone (eds.), *Teaching and Language Corpora*, 100–115. Axon: Routledge.

Jolliffe, I. T. 2002. *Principal Component Analysis* (Springer Series in Statistics). 2nd ed. New York: Springer.

Kachru, Yamuna & Larry E. Smith. 2008. *Cultures, contexts and world Englishes* (ESL & Applied Linguistics Professional Series). New York: Routledge.

Kaiser, Henry F & John Rice. 1974. Little Jiffy, Mark IV. *Educational and psychological measurement*. Sage Publications Sage CA: Thousand Oaks, CA 34(1). 111–117.

Kaszubski, Przemek. 1998. Enhancing a writing textbook: a national perspective. In *Symposium proceedings of the First International Symposium o n Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 172–185. The Chinese University of Hong Kong: Hong Kong.

Kerres, Michael. 2020. Against All Odds: Education in Germany Coping with Covid-19. *Postdigital Science and Education* 2(3). 690–694. https://doi.org/10.1007/s42438-020-00130-7.

Kheovichai, Baramee. 2017. A Corpus-Based Analysis of BE + Being + Adjective in English from the Appraisal Framework Perspective. *Silpakorn University Journal of Social Sciences, Humanities, and Arts* 17(2). 97–132.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography* 1(1). 7–36.

Kim, Daejin & Joan Kelly Hall. 2002. The role of an interactive book reading program in the development of second language pragmatic competence. *The Modern Language Journal* 86(3). 332–348.

Knowles, Jared E. & Carl Frederick. 2020. *merTools: Tools for analyzing mixed effect regression models*. Manual. https://CRAN.R-project.org/package=merTools (12 July, 2020).

Koch, Peter & Wulf Oesterreicher. 1985. Language of Immediacy - Language of Distance: Orality and Literacy from the Perspective of Language Theory and Linguistic History. In Claudia Lange, Beatrix Weber & Göran Wolf (eds.), *Communicative Spaces. Variation, Contact, and Change: Papers in Honour of Ursula Schaefer*, 441–473. Frankfurt am Main: Peter Lang.

Koch, Peter & Wulf Oesterreicher. 2011. *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch* (Romanistische Arbeitshefte 31). 2., aktualisierte und erw. Aufl. Berlin ; New York: De Gruyter.

Koprowski, M. 2005. Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal* 59(4). 322–332. https://doi.org/10.1093/elt/cci061.

Koprowski, Mark. 2005. Investigating the Usefulness of Lexical Phrases in Contemporary Coursebooks. *ELT Journal* 59(4). 322–332. https://doi.org/10.1093/elt/cci061.

Kosmas, Panagiotis, Antigoni Parmaxi, Maria Perifanou & Anastasios A. Economides. 2021. Open Educational Resources for Language Education: Towards the Development of an e-Toolkit. In Panayiotis Zaphiris & Andri Ioannou (eds.), *Learning and Collaboration Technologies: New Challenges and Learning Experiences. 8th International Conference, LCT 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I* (Lecture Notes in Computer Science), 65–79. Cham: Springer International Publishing.

Krashen, Stephen. 1982. *Principles and Practice in Second Language Acquisition* (Language Teaching Methodology Series). Oxford; New York: Pergamon.

Krashen, Stephen. 1985. *Second language acquisition and second language learning* (Language Teaching Methodology Series). Reprinted. Oxford: Pergamon Press.

Kuckartz, Udo. 2014. *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung.* Weinheim: Beltz Juventa.

Kuhl, Patricia K. & Andrew N. Meltzoff. 1996. Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America* 100(401). 2425–2438.

Kultusministerkonferenz. 2003. Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Mittleren Schulabschluss. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2003/2003_12_04-BS-erste-Fremdsprache.pdf (21 January, 2017).

Kultusministerkonferenz. 2012. Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch) für die Allgemeine Hochschulreife. Beschluss der Kultusministerkonferenz vom 18.10.2012. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. http://www.kmk.org/bildung-schule/qualitaetssicherung-in-schulen/bildungsstandards/dokumente.html (26 November, 2018).

Kurtz, Jürgen. 2019. Lehrwerkgestütztes Fremdsprachenlernen im digitalen Wandel. In Eva Burwitz-Melzer, Claudia Riemer & Lars Schmelter (eds.), *Das Lehren und Lernen von Fremd- und Zweitsprachen im digitalen Wandel: Arbeitspapiere der 39. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts* (Gießener Beiträge zur Fremdsprachendidaktik), 114–125. Tübingen: Narr Francke Attempto.

Lakoff, George & Mark Johnson. 2003. *Metaphors we live by.* Chicago: University of Chicago Press.

Landesregierung Nordrhein-Westfalen. 2016. Lernen im digitalen Wandel. Unser Leitbild 2020 für Bildung in Zeiten der Digitalisierung. https://www.land.nrw/sites/default/files/asset/document/leitbild_lernen_im_digitalen_wandel.pdf (22 January, 2022).

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar: Theoretical prerequisites.* Stanford University Press.

Langacker, Ronald W. 2001. Cognitive linguistics, language pedagogy, and the English present tense. In Martin Pütz, Susanne Niemeier & René Dirven (eds.), *Applied Cognitive Linguistics. I: Theory and Language Acquisition:*, 3–39. Berlin; New York: De Gruyter. https://doi.org/10.1515/9783110866247.

Langacker, Ronald W. 2008. *Cognitive Grammar: A Basic Introduction.* Oxford: Oxford University Press.

Larsen-Freeman, Diane. 1997. Chaos/complexity science and second language acquisition. *Applied linguistics.* Oxford University Press 18(2). 141–165.

Larsen-Freeman, Diane. 2006. The Emergence of Complexity, Fluency, and Accuracy in the Oral and Written Production of Five Chinese Learners of English. *Applied Linguistics* 27(4). 590–619. https://doi.org/10.1093/applin/aml029.

Larsen-Freeman, Diane. 2018. Looking ahead: Future directions in, and future research into, second language acquisition. *Foreign Language Annals* 51(1). 55–72. https://doi.org/10.1111/flan.12314.

Lavi-Rotbain, Ori & Inbal Arnon. 2022. The learnability consequences of Zipfian distributions in language. *Cognition* 223. 105038. https://doi.org/10.1016/j.cognition.2022.105038.

Le Foll, Elen. 2016. *Collocational competence in English as a B language. A case study of potential conference interpreting students.* Technische Hochschule Köln Unpublished M.A. thesis.

Le Foll, Elen. 2017. Textbook English: A corpus-based analysis of language use in German and French EFL textbooks. Presented at the Corpus Linguistics 2017 Conference, University of Birmingham. https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper230.pdf (20 October, 2017).

Le Foll, Elen. 2018a. "They were walking in a corridor when suddenly the mummy appeared." A Corpus-based Study of Narrative Texts in Secondary School EFL Textbooks. Presented at the 13th Teaching Language and Corpora (TaLC) Conference, Cambridge, UK. https://doi.org/10.13140/RG.2.2.31141.81124. http://rgdoi.net/10.13140/RG.2.2.31141.81124 (8 February, 2022).

Le Foll, Elen. 2018b. Raising, *discovering and *exploiting awareness: Using the Internet as a source of collocational knowledge. Presented at the OLLReN Annual Online Conference: Research into using Technology for Language Learning. https://www.researchgate.net/publication/328368290_Raising_discovering_and_exploiting_awareness_Using_the_Internet_as_a_source_of_collocational_knowledge.

Le Foll, Elen. 2020a. Exploring the Registers of School EFL Textbooks Using Multi-Dimensional Analysis. Presented at the Paper presented at the ICAME41, Heidelberg University.

Le Foll, Elen. 2020b. Development and Evaluation of a Corpus Linguistics Seminar in Pre-Service Teacher Training. Presented at the Teaching and Language Corpora Conference (TaLC) 2020, Perpignan.

Le Foll, Elen. 2020c. Issues in Compiling and Exploiting Textbook Corpora. Presented at the Japanese Association for English Corpus Studies 2020, Tokyo. https://doi.org/10.13140/RG.2.2.32006.60487.

Le Foll, Elen. 2021a. *A New Tagger for the Multi-Dimensional Analysis of Register Variation in English.* Osnabrück University: Institute of Cognitive Science Unpublished M.Sc. thesis.

Le Foll, Elen. 2021b. Register Variation in School EFL Textbooks. *Register Studies* 3(2). https://doi.org/10.1075/rs.20009.lef.

Le Foll, Elen. 2021c. *Creating Corpus-Informed Materials for the English as a Foreign Language Classroom: A step-by-step guide for (trainee) teachers using online resources* (Open Educational Resource). 3rd edn. https://elenlefoll.pressbooks.com (30 July, 2021).

Le Foll, Elen. 2021d. Learning by doing: ein Online-Projektseminar zur Vermittlung von digitalen Medien- und Datenkompetenzen für den Fremdsprachenunterricht. Presented at the DigiRom - Fremdsprachliche Lehrer*innenbildung digital?, Osnabrück University.

Le Foll, Elen. 2021e. *Introducing the Multi-Feature Tagger of English (MFTE).* Perl. Osnabrück University. https://github.com/elenlefoll/MultiFeatureTaggerEnglish (5 January, 2022).

Le Foll, Elen. forthcoming. Textbooks and Corpus Linguistics: the case of causative constructions. In Kieran Harrington & Patricia R. Ronan (eds.), *Corpus Linguistics*

in the English Language Teaching Classroom - Research and Practice. Palgrave MacMillan.

Lê, Sébastien, Julie Josse & François Husson. 2008. FactoMineR : An R Package for Multivariate Analysis. *Journal of Statistical Software* 25(1). https://doi.org/10.18637/jss.v025.i01.

Leacock, Claudia, Martin Chodorow, Michael Gamon & Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners* (Synthesis Lectures on Human Language Technologies). Vol. 10. San Rafael: Morgan & Claypool. https://doi.org/10.2200/S00275ED1V01Y201006HLT009.

Lee, David YW. 2001. Genres, Registers, Text Types, Domain, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning & Technology* 5(3). 37–72.

Lee, Hansol, Mark Warschauer & Jang Ho Lee. 2019. The Effects of Corpus Use on Second Language Vocabulary Learning: A Multilevel Meta-analysis. *Applied Linguistics* 40(5). 721–753. https://doi.org/10.1093/applin/amy012.

Lee, Yong Wey David. 2000. *Modelling variation in spoken and written language : the multi-dimensional approach revisited*. Lancaster: Lancaster University Unpublished PhD thesis.

Leech, G, P Rayson & A Wilson. 2001. *Word Frequencies in Written and Spoken English*. Harlow: Longman.

Leńko-Szymańska, Agnieszka. 2017. Training teachers in data-driven learning: Tackling the challenge. *Language Learning & Technology* 21(3). 217–241.

Leńko-Szymańska, Agnieszka & Alex Boulton. 2015. Introduction: Data-driven learning in language pedagogy. In Agnieszka Leńko-Szymańska & Alex Boulton (eds.), *Studies in Corpus Linguistics*, vol. 69, 1–14. Amsterdam: John Benjamins.

Lennon, Paul. 1996. Getting 'easy' verbs wrong at the advanced level. *IRAL - International Review of Applied Linguistics in Language Teaching* 34(1). 23–36.

Lenth, Russell. 2020. *emmeans: Estimated marginal means, aka least-squares means*. Manual. https://CRAN.R-project.org/package=emmeans (12 July, 2020).

Leroy, Michel. 2012. *Les manuels scolaires : situation et perspectives*. Inspection générale de l'éducation nationale. https://www.education.gouv.fr/les-manuels-scolaires-situation-et-perspectives-6017 (1 August, 2018).

Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam; Philadelphia, PA: John Benjamins.

Lewis, Michael. 1993. *The Lexical Approach*. Hove: Language Teaching Publications.

Lewis, Michael. 1997. *Implementing the Lexical Approach: Putting Theory into Practice*. Hove: Language Teaching Publications.

Lewis, Michael. 2009. *The lexical approach: the state of ELT and a way forward*. 6. [print.]. London: Heinle.

Liedke, Martina. 2013. Mit Transkripten Deutsch lernen. In Sandro M. Moraldo & Federica Missaglia (eds.), *Gesprochene Sprache im DaF-Unterricht: Grundlagen – Ansätze – Praxis*, 243–266. Heidelberg: Winter.

Lier, Leo van. 2013. *Interaction in the Language Curriculum: Awareness, Autonomy and Authenticity*. Hoboken: Taylor and Francis.

Limberg, Holger. 2016. "Always remember to say Please and Thank You": Teaching Politeness with German EFL Textbooks. *Pragmatics & Language Learning* 265.

Little, David, Sean Devitt & David Singleton. 1989. *Learning foreign languages from authentic texts: Theory and practice*. Dublin: Authentik.

Little, David, Seán Devitt & David Singleton. 2002. The communicative approach and authentic texts. In *Teaching modern languages*, 51–55. London: Routledge.

Littlejohn, Andrew. 2011. The analysis of language teaching materials: Inside the Trojan Horse. In Brian Tomlinson (ed.), *Materials development in language teaching*, 179–211. Cambridge: Cambridge University Press.

Liu, Dilin. 2011. The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *Tesol Quarterly* 45(4). 661–688.

Liu, Eric T.K. & Philip M. Shaw. 2001. Investigating learner vocabulary: A possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *International Review of Applied Linguistics in Language Teaching* 39(3). 171–194. https://doi.org/10.1515/iral.2001.001.

Ljung, Magnus. 1990. *A study of TEFL vocabulary* (Stockholm Studies in English 78). Stockholm: Almqvist & Wiksell International.

Ljung, Magnus. 1991. Swedish TEFL meets reality. In Stig Johansson & Anna-Brita Stenström (eds.), *English Computer Corpora: Selected Papers and Research Guide*, 245–256.

Loewen, Shawn & Talip Gonulal. 2015. Exploratory Factor Analysis and Principal Components Analysis. In Luke Plonsky (ed.), *Advancing quantitative methods in second language research* (Second Language Acquisition Research Series), 182–211. New York: Routledge.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014. *International Journal of Corpus Linguistics* 22(3). 319–344. https://doi.org/10.1075/ijcl.22.3.02lov.

Love, Robbie, Abi Hawtin & Andrew Hardie. 2018. The British National Corpus 2014: User manual and reference guide. http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf (15 November, 2018).

Lüdecke, Daniel. 2020. *sjPlot: Data visualization for statistics in social science*. Manual. https://CRAN.R-project.org/package=sjPlot (12 July, 2020).

Luke, Steven G. 2017. Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods* 49(4). 1494–1502. https://doi.org/10.3758/s13428-016-0809-y.

MacWhinney, B. 2006. Emergentism--Use Often and With Care. *Applied Linguistics* 27(4). 729–740. https://doi.org/10.1093/applin/aml035.

Madlener, Karin. 2018. Do findings from artificial language learning generalize to second language classrooms? In Andrea Tyler, Lourdes Ortega & Mariko Uno (eds.), *Usage-inspired L2 instruction: researched pedagogy* (Language Learning & Language Teaching 49), 211–234. Amsterdam: John Benjamins.

Maechler, Martin. 2019. *Rmpfr: R MPFR - Multiple Precision Floating-Point Reliable*. https://CRAN.R-project.org/package=Rmpfr.

Mair, Christian. 2012. Progressive and Continuous Aspect. In Robert Binnick (ed.), *The Oxford Handbook of Tense and Aspect*, 803–827. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195381979.013.0028.

Mair, Christian & Marianne Hundt. 1995. Why is the Progressive Becoming More Frequent in English? A Corpus-Based Investigation of Language Change in Progress. *Zeitschrift für Anglistik und Amerikanistik* 43. 111–122.

Manning, Christopher D. 2011. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In Alexander F. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing* (Lecture Notes in Computer Science), vol. 6608, 171–189. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-19400-9_14.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, Mass: MIT Press.

Martin, James R. & David Rose. 2008. *Genre relations: mapping culture* (Equinox Textbooks and Surveys in Linguistics). London; Oakville, CT: Equinox Pub.

Martin, J.R. 2009. Genre and language learning: A social semiotic perspective. *Linguistics and Education* 20(1). 10–21. https://doi.org/10.1016/j.linged.2009.01.003.

Martinez, Ron & Norbert Schmitt. 2012. A Phrasal Expressions List. *Applied Linguistics* 33(3). 299–320. https://doi.org/10.1093/applin/ams010.

Matthiessen, Christian M.I.M. 2019. Register in Systemic Functional Linguistics. *Register Studies* 1(1). 10–41. https://doi.org/10.1075/rs.18010.mat.

Mauranen, Anna. 2003. The Corpus of English as Lingua Franca in Academic Settings. *TESOL Quarterly* 37(3). 513–27.

Mauranen, Anna. 2004a. Speech corpora in the classroom. In Guy Aston, Silvia Bernardini & Dominic Stewart (eds.), *Corpora and language learners* (Studies in Corpus Linguistics v. 17), 195–211. Amsterdam: John Benjamins.

Mauranen, Anna. 2004b. Spoken corpus for an ordinary learner. In John McHardy Sinclair (ed.), *How to Use Corpora in Language Teaching* (Studies in Corpus Linguistics), 89–108. Amsterdam; Philadelphia: John Benjamins.

Mauranen, Anna, Niina Hynninen & Elina Ranta. 2010. English as an academic lingua franca: The ELFA project. *English for Specific Purposes* 29(3). 183–190. https://doi.org/10.1016/j.esp.2009.10.001.

Mayring, Philipp. 2010. *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (Beltz Pädagogik). Weinheim: Beltz.

McCarthy, Michael. 1998. *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.

McCarthy, Michael & Ronald Carter. 1995. Spoken grammar: what is it and how can we teach it? *ELT journal* 49(3). 207–218.

McCarthy, Michael & Ronald Carter. 2001. Size isn't everything: spoken English, corpus, and the classroom. *Tesol Quarterly* 35(2). 337–340.

McEnery, Tony & Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395.

McEnery, Tony & Richard Xiao. 2011. What Corpora Can Offer in Language Teaching and Learning. In Eli Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning* (ESL & Applied Linguistics Professional Series), 364–380. New York, NY: Routledge.

McManus, Kevin. 2021. Are Replication Studies Infrequent Because of Negative Attitudes? Insights from a Survey of Attitudes and Practices in Second Language Research. *Studies in Second Language Acquisition* 1–14. https://doi.org/10.1017/S0272263121000838.

Meriläinen, Lea. 2018. The progressive form and its functions in spoken learner English: Tracing the effects of an exposure-rich learning environment. *International Journal of Learner Corpus Research* 4(2). 164–194. https://doi.org/10.1075/ijlcr.17002.mer.

Meunier, Fanny & Céline Gouverneur. 2007. The treatment of phraseology in ELT textbooks. In Encarnación Hidalgo, Luis Quereda & Juan Santana (eds.), *Corpora in the Foreign Language Classroom*, 119–139. Amsterdam, New York: Brill Rodopi. https://doi.org/10.1163/9789401203906_009.

Meunier, Fanny & Céline Gouverneur. 2009. New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. In Karin Aijmer (ed.), *Studies in Corpus Linguistics*, vol. 33, 179–201. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.33.16meu.

Meunier, Fanny & Randi Reppen. 2015. Corpus versus non-corpus-informed pedagogical materials: grammar as the focus. In Douglas Biber & Randi Reppen (eds.), *The Cambridge handbook of English corpus linguistics* (Cambridge Handbooks in Language and Linguistics), 498–514. Cambridge: Cambridge University Press.

Meurers, Detmar. 2015. Learner corpora and natural language processing. In Sylviane Granger, Gaetanelle Gilquin & Fanny Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, 537–566. Cambridge: Cambridge University Press.

Meurers, Detmar, Kordula De Kuthy, Florian Nuxoll, Björn Rudzewitz & Ramon Ziai. 2019. Scaling Up Intervention Studies to Investigate Real-Life Foreign Language Learning in School. *Annual Review of Applied Linguistics* 39. 161–188. https://doi.org/10.1017/S0267190519000126.

Meurers, Detmar, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf & Niels Ott. 2010. Enhancing Authentic Web Pages for Language Learners. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications*, 9. Los Angeles. http://www.sfs.uni-tuebingen.de/~dm/papers/meurers-ziai-et-al-10.pdf (21 January, 2022).

Miekley, Joshua. 2005. ESL textbook evaluation checklist. *The Reading Matrix* 5(2).

Miller, Don. 2011. ESL Reading Textbooks vs. University Textbooks: Are We Giving Our Students the Input They May Need? *Journal of English for Academic Purposes* 10(1). 32–46. https://doi.org/10.1016/j.jeap.2010.12.002.

Milroy, James & Lesley Milroy. 2012. *Authority in language: investigating standard English* (Routledge Linguistics Classics). Abingdon, Oxon; New York: Routledge.

Mindt, Dieter. 1987. *Sprache, Grammatik, Unterrichtsgrammatik: futurischer Zeitbezug im Englischen* (Schule und Forschung). Frankfurt am Main: Diesterweg.

Mindt, Dieter. 1992. *Zeitbezug Im Englischen: Eine Didaktische Grammatik Des Englischen Futurs* (Tübinger Beiträge Zur Linguistik). Tübingen: Gunter Narr Verlag.

Mindt, Dieter. 1995a. *An Empirical Grammar of the English verb: Modal Verbs*. Berlin: Cornelsen.

Mindt, Dieter. 1995b. Schulgrammatik vs. Grammatik der englischen Sprache. In Claus Gnutzmann & Frank G. Königs (eds.), *Perspektiven des Grammatikunterrichts*, 47–68. Tübingen: Gunter Narr Verlag.

Mindt, Dieter. 1996. English corpus linguistics and the foreign language teaching syllabus. In J. Thomas & M. Short (eds.), *Using Corpora for Language Research*, 232–247. Harlow: Longmann.

Mindt, Dieter. 1997a. Corpora and the Teaching of English in Germany. In Anne Wichmann (ed.), *Teaching and language corpora* (Applied Linguistics and Language Study), 40–50. 1. publ. London: Longman.

Mindt, Dieter. 1997b. Complementary distribution, gradience and overlap in corpora and in ELT: Analysing and teaching the progressive. In *From Ælfric to the New York Times. Studies in English Corpus Linguistics Age and Computers*, vol. 19, 227–238. Amsterdam: Rodopi.

Mishan, F. 2005. *Designing Authenticity Into Language Learning Materials* (Play Text Series). Portland, Oregon: Intellect Books.

Mishra, Punya & Matthew J. Koehler. 2006. Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record* 108(6). 1017–1054.

Mohr, Susanne, Sandra Jansen & Julia Forsberg. 2019. European English in the EFL classroom?: Teacher attitudes towards target varieties of English in Sweden and Germany. *English Today* 1–7. https://doi.org/10.1017/S0266078419000403.

Möller, Stefan. 2016. Sourcebook Rather Than Coursebook. Lernerorientiert mit dem Lehrwerk arbeiten. *Der fremdsprachliche Unterricht Englisch* 50(143). 12–18.

Möller, Verena. 2017. A statistical analysis of learner corpus data, experimental data and individual differences: Monofactorial vs. multifactorial approaches. In Pieter de Haan, Rina de Vries & Sanne van Vuuren (eds.), *Language, learners and levels: Progression and variation*, 409–439. Louvain-la-Neuve: Presses Universitaires de Louvain.

Möller, Verena. 2020. From pedagogical input to learner output: Conditionals in EFL and CLIL teaching materials and learner language. *Pedagogical Linguistics* 1(2). 95–124. https://doi.org/10.1075/pl.00001.mol.

Moreno, Ana I. 2003. Matching Theoretical Descriptions of Discourse and Practical Applications to Teaching: The Case of Causal Metatext. *English for Specific Purposes* 22(3). 265–295. https://doi.org/10.1016/S0889-4906(02)00021-2.

Moyer, Alene. 2008. Input as a Critical Means to an End: Quantity and Quality of Experience in L2 Phonological Attainment. In Thorsten Piske & Martha Young-Scholten (eds.), *Input Matters in SLA*, 159–174. Multilingual Matters. https://doi.org/10.21832/9781847691118-011.

Moyer, Alene. 2013. *Foreign Accent: The Phenomenon of Non-Native Speech*. Cambridge: Cambridge University Press.

Mukherjee, Joybrato. 2004. Bridging the Gap between Applied Corpus Linguistics and the Reality of English Language Teaching in Germany. In U. Connor & T. Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*, 239–250. Amsterdam: Rodopi.

Mukundan, Jayakaran. 2010. Retrotext- E 1.0: The Beginnings of Computer-based ELT Textbook Evaluation. *Advances in Language and Literary Studies* 1(2). 270–280. https://doi.org/10.7575/aiac.alls.v.1n.2p.270.

Mukundan, Jayakaran & Touran Ahour. 2010. A review of Textbook Evaluation Checklists across Four Decades (1970–2008). In Brian Tomlinson & Hitomi Masuhara (eds.), *Research for Materials Development in Language Learning: Evidence for Best Practice*, 336–352. London: Bloomsbury.

Mukundan, Jayakaran, Amelia Leong Chiew Har & Vahid Nimehchisalem. 2012. Distribution of Articles in Malaysian Secondary School English Language

Textbooks. *English Language and Literature Studies* 2(2). 62–70. https://doi.org/10.5539/ells.v2n2p62.

Mukundan, Jayakaran & Norwati Roslim. 2009. Textbook Representation of Prepositions. *English Language Teaching* 2(4). 13–24. https://doi.org/10.5539/elt.v2n4p13.

Müller, Simone. 2005. *Discourse Markers in Native and Non-native English Discourse* (Pragmatics & Beyond New Series). Vol. 138. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.138.

Mundry, Roger & Christina Sommer. 2007. Discriminant function analysis with nonindependent data: consequences and an alternative. *Animal Behaviour* 74(4). 965–976. https://doi.org/10.1016/j.anbehav.2006.12.028.

Munro, Murray J. & Tracey M. Derwing. 1998. The Effects of Speaking Rate on Listener Evaluations of Native and Foreign-Accented Speech. *Language Learning* 48(2). 159–182. https://doi.org/10.1111/1467-9922.00038.

Murphy, M. Lynne. 2003. *Semantic Relations and the Lexicon: Antonymy, Synonymy and other Paradigms*. Cambridge: Cambridge University Press.

Murphy, Raymond, Brigit Viney, Miles Craven & Diane Cranz. 2009. *English grammar in use: a self-study reference and practice book for intermediate students of English ; with answers*. 3. ed., 14. print. Cambridge: Cambridge University Press.

Nakagawa, Shinichi, Paul C. D. Johnson & Holger Schielzeth. 2017. The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface* 14(134). 20170213. https://doi.org/10.1098/rsif.2017.0213.

Nelson, Mike. 2022. Corpus-based evaluation of English language learning textbooks. In Reka R. Jablonkai & Eniko Csomay (eds.), *The Routledge Handbook of Corpora and English Language Teaching and Learning* (Routledge Handbooks in Applied Linguistics). London: Routledge.

Nesselhauf, Nadja. 2004. How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In Guy Aston, Silvia Bernardini & Dominic Stewart (eds.), *Corpora and Language Learners* (Studies in Corpus Linguistics), vol. 17, 109–124. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.17.08nes.

Nesselhauf, Nadja. 2005. *Collocations in a Learner Corpus*. Amsterdam; Philadelphia: John Benjamins.

Nesselhauf, Nadja & Ute Römer. 2007. Lexical-grammatical patterns in spoken English: The case of the progressive with future time reference. *International journal of corpus linguistics* 12(3). 297–333.

Neumann, Stella & Stephanie Evert. 2021. A register variation perspective on varieties of English. In Elena Seoane & Douglas Biber (eds.), *Corpus-based approaches to register variation* (Studies in Corpus Linguistics volume 103), 144–178. Amsterdam; Philadelphia: Benjamins.

Niemeier, Susanne. 2017. *Task-based grammar teaching of English: Where cognitive grammar and task-based language teaching meet*. Tübingen: Narr Francke Attempto Verlag.

Nini, Andrea. 2014. *Multidimensional Analysis Tagger (MAT)*. http://sites.google.com/site/multidimensionaltagger (18 September, 2019).

Nini, Andrea. 2019. The Multi-Dimensional Analysis Tagger. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 67–96. New York: Bloomsbury.

Ninio, Anat. 1999. Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of child language*. Cambridge University Press 26(3). 619–653. https://doi.org/10.1017/s0305000999003931.

Nold, Günter. 1998. Die Arbeit mit dem Lehrwerk. In Johannes P. Timm (ed.), *English lernen und lehren – Didaktik des Englischunterrichts*, 127–136. Berlin: Cornelsen.

Nordlund, Marie. 2016. EFL Textbooks for Young Learners: A Comparative Analysis of Vocabulary. *Education Inquiry* 7(1). 47–68. https://doi.org/10.3402/edui.v7.27764.

Novakova, Iva & Dirk Siepmann (eds.). 2020. *Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-23744-8.

Oelkers, Jürgen. 2008. Lehrplanentwicklung, Lehrmittel und Bildungsstandards. Presented at the Klausurtagung der Leitungskonferenz des Staatsinstituts für Schulentwicklung und Bildungsforschung, St. Quirin. https://www.ife.uzh.ch/research/emeriti/oelkersjuergen/vortraegeprofoelkers/vortraege2008/339_StQuirin.pdf (8 February, 2022).

O'Keeffe, Anne, Michael McCarthy & Ronald Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Papert, Seymour A. 2020. *Mindstorms: Children, Computers, And Powerful Ideas*. New York: Hachette.

Paquot, Magali & Luke Plonsky. 2017. Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3(1). 61–94. https://doi.org/10.1075/ijlcr.3.1.03paq.

Peacock, Matthew. 1997. The effect of authentic materials on the motivation of EFL learners. *ELT journal* 51(2). 144–156.

Pedersen, Ted. 1996. Fishing for Exactness. *arXiv:cmp-lg/9608010*. http://arxiv.org/abs/cmp-lg/9608010 (11 February, 2020).

Pérez-Paredes, Pascual & Malena Abad. forthcoming. Integrating language teachers' voices in the design and exploitation of Spanish corpora in the UK. In Henry Tyne & Stefania Spina (eds.), *Applying corpora in teaching and learning Romance languages*. Amsterdam: John Benjamins.

Pérez-Paredes, Pascual & Geraldine Mark (eds.). 2021. *Beyond Concordance Lines: Corpora in Language Education* (Studies in Corpus Linguistics 102). Amsterdam; Philadelphia: John Benjamins.

Pérez-Paredes, Pascual, Geraldine Mark & Anne O'Keeffe. 2020. *The impact of usage-based approaches on second language learning and teaching*. Cambridge Education Research Reports. Cambridge: Cambridge University Press. https://www.cambridge.org/partnership/research/impact-usage-based-approaches-second-language-learning-and-teaching (9 July, 2021).

Phakiti, Aek & Luke Plonsky. 2018. Reconciling Beliefs about L2 Learning with SLA Theory and Research. *RELC Journal*. SAGE Publications Ltd 49(2). 217–237. https://doi.org/10.1177/0033688218781970.

Piaget, Jean. 2013. *The Construction Of Reality In The Child*. Oxon: Taylor & Francis.

Picoral, Adriana, Shelley Staples & Randi Reppen. 2021. Automated annotation of learner English: An evaluation of software tools. *International Journal of Learner Corpus Research* 7(1). 17–52. https://doi.org/10.1075/ijlcr.20003.pic.

Pienemann, Manfred. 1984. Psychological constraints on the teachability of languages. *Studies in second language acquisition* 6(2). 186–214.

Popescu, Marius. 2011. Studying Translationese at the Character Level. In *Proceedings of Recent Advances in Natural Language Processing*, 634–639. Hissar, Bulgaria.

Popper, Karl R. 1959. *The logic of scientific discovery*. New York: Basic Books.

Posner, Michael I. & Steven W. Keele. 1968. On the genesis of abstract ideas. *Journal of Experimental Psychology* 77(3, Pt.1). 353–363. https://doi.org/10.1037/h0025953.

Posner, Michael I. & Steven W. Keele. 1970. Retention of abstract ideas. *Journal of Experimental Psychology*. American Psychological Association 83(2, Pt. 1). 304–308.

Prabhu, Nagore Seshagiri. 1989. Materials as Support; Materials as Constraint. *Guidelines: A Periodical for Classroom Language Teachers*. ERIC 11(1). 66–74.

Prodromou, Luke. 1992. What culture? Which culture? Cross-cultural factors in language learning. *ELT journal* 46(1). 39–50.

Prodromou, Luke. 1996. Correspondence. *ELT Journal* 50(1). 88–89. https://doi.org/10.1093/elt/50.1.88.

Prodromou, Luke. 2003. In search of the successful user of English How a corpus of non-native speaker language could impact on EFL teaching. *Modern English Teacher* 12(2). 5–15.

Prowse, Philip. 1998. How writers write: testimony from authors. In Brian Tomlinson (ed.), *Materials Development in Language Teaching*, 130–145. Cambridge: Cambridge University Press.

Quaglio, Paulo. 2009. *Television Dialogue: The sitcom* Friends *vs. natural conversation* (Studies in Corpus Linguistics 36). Amsterdam: John Benjamins.

Quetz, Jürgen. 1999. Welche Normen muss ein Lehrwerk erfüllen? In Klaus Vogel & Wolfgang Börner (eds.), *Lehrwerke im Fremdsprachenunterricht: lernbezogene, interkulturelle und mediale Aspekte* (Fremdsprachen in Lehre und Forschung 23), 10–24. Bochum: AKS-Verl.

Quirk, Randolph. 1995. *Grammatical and Lexical Variance in English*. London: Routledge.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London; New York: Longman.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Manual. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/ (29 January, 2021).

Ramjattan, Vijay A. 2019. The white native speaker and inequality regimes in the private English language school. *Intercultural Education* 30(2). 126–140. https://doi.org/10.1080/14675986.2018.1538043.

Ranalli, James M. 2003. *The treatment of key vocabulary learning strategies in current ELT course books: repetition, resource use, recording*. U.K.: University of Birmingham M.A. thesis. http://www.birmingham.ac.uk/documents/college-artslaw/cels/essays/matefltesldissertations/ranallidiss.pdf (10 August, 2017).

Rankin, Tom. 2010. Advanced Learner Corpus Data and Grammar Teaching: Adverb Placement. In Mari Carmen Campoy Cubillo, Begoña Bellés-Fortuño & Maria

Lluïsa Gea-Valor (eds.), *Corpus-Based Approaches to English Language Teaching*, 205–2015. London ; New York: Continuum International Publishing Group.

Rasch, Björn, Malte Friese, Wilhelm Hofmann & Ewald Naumann. 2014. Einfaktorielle Varianzanalyse. In Björn Rasch, Malte Friese, Wilhelm Hofmann & Ewald Naumann (eds.), *Quantitative Methoden 2: Einführung in die Statistik für Psychologen und Sozialwissenschaftler* (Springer-Lehrbuch), 1–34. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-43548-9_1. https://doi.org/10.1007/978-3-662-43548-9_1 (6 January, 2021).

Rautionaho, Paula. 2014. *Variation in the progressive a corpus-based study into world Englishes*. Tampere: Tampere University Press PhD thesis. https://tampub.uta.fi/bitstream/handle/10024/96287/978-951-44-9636-3.pdf?sequence=1 (14 June, 2019).

Rautionaho, Paula & Sandra C. Deshors. 2018. Progressive or not progressive?: Modeling the constructional choices of EFL and ESL writers. *International Journal of Learner Corpus Research* 4(2). 225–252. https://doi.org/10.1075/ijlcr.16019.rau.

Rautionaho, Paula, Sandra C. Deshors & Lea Meriläinen. 2018. Revisiting the ENL-ESL-EFL continuum: A multifactorial approach to grammatical aspect in spoken Englishes. *ICAME Journal* 42(1). 41–78. https://doi.org/10.1515/icame-2018-0004.

Rayson, Paul. 2018. USAS English Wmatrix3 web tagger. http://ucrel-api.lancaster.ac.uk/usas/tagger.html (7 December, 2019).

Reda, Ghsoon. 2003. English Coursebooks: Prototype Texts and Basic Vocabulary Norms. *ELT Journal* 57(3). 260–268. https://doi.org/10.1093/elt/57.3.260.

Reder, Stephen, Kathryn Harris & Kristen Setzler. 2003. The multimedia adult ESL learner corpus. *TESOL Quarterly*. JSTOR 37(3). 546–557.

Reinders, Hayo & Rod Ellis. 2009. The Effects of Two Types of Input on Intake and the Acquisition of Implicit and Explicit Knowledge. In Rod Ellis, Shawn Loewen, Catherine Elder, Hayo Reinders, Rosemary Erlam & Jenefer Philp (eds.), *Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching*, 281–302. Bristol: Multilingual Matters.

Rencher, Alvin C. 1992. Interpretation of Canonical Discriminant Functions, Canonical Variates, and Principal Components. *American Statistical Association* 46(3). 217–225.

Reppen, Randi. 1994a. *Variation in Elementary Student Writing*. Northern Arizona University Unpublished Ph.D. Dissertation.

Reppen, Randi. 1994b. *Variation in elementary student writing*. Northern Arizona University Unpublished Ph. D. Dissertation.

Reppen, Randi. 2013. Register variation in student and adult speech and writing. In Susan Conrad & Douglas Biber (eds.), *Variation in English: Multi-dimensional studies*, 187–199.

Revelle, William. 2020. *psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois. https://CRAN.R-project.org/package=psych (20 April, 2021).

Révész, Andrea & Tineke Brunfaut. 2013. Text Characteristics of Task Input and Difficulty in Second Language Listening Comprehension. *Studies in Second Language Acquisition* 35(1). 31–65. https://doi.org/10.1017/S0272263112000678.

Richards, Jack C. 2001. *Curriculum Development in Language Teaching*. 1st edn. Cambridge University Press. https://doi.org/10.1017/CBO9780511667220.

Richards, Jack C. 2015. *Key Issues in Language Teaching*. Cambridge University Press.

Richardson, Leonard. 2015. Beautiful Soup 4.4.0 documentation. https://beautiful-soup-4.readthedocs.io/en/latest/ (7 January, 2022).

Rinker, Tyler W. 2018. *textstem: Tools for stemming and lemmatizing text in R*. Buffalo, New York. http://github.com/trinker/textstem (1 February, 2019).

Rinvolucri, Mario. 1999. The UK, EFLese sub-culture and dialect. *Folio* 5(2). 12–14.

Roberts, Celia & Melanie Cooke. 2009. Authenticity in the Adult ESOL Classroom and Beyond. *TESOL Quarterly* 43(4). 620–642. https://doi.org/10.1002/j.1545-7249.2009.tb00189.x.

Robinson, Peter J. & Nick C. Ellis (eds.). 2008. *Handbook of cognitive linguistics and second language acquisition*. New York: Routledge.

Rohde, Ada Ragna. 2001. *Analyzing PATH: The interplay of verbs, prepositions and constructional semantics*. Houston, TX: Rice University.

Rohleder, Bernhard. 2019. Smart School - Auf dem Weg zur digitalen Schule. Presented at the Bitkom Research GmbH, Berlin. https://www.bitkom.org/sites/default/files/2019-03/Pr%C3%A4sentation%20Bitkom-PK%20Bildungskonferenz%2012.03.2019_final.pdf (29 September, 2020).

Römer, Ute. 2004a. Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In Guy Aston, Silvia Bernardini & Dominic Stewart (eds.), *Studies in Corpus Linguistics*, vol. 17, 151–168. Amsterdam: John Benjamins.

Römer, Ute. 2004b. A Corpus-Driven Approach to Modal Auxiliaries and Their Didactics. In John McH. Sinclair (ed.), *How to Use Corpora in Language Teaching* (Studies in Corpus Linguistics), vol. 12, 185–199. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.12.14rom.

Römer, Ute. 2005. *Progressives, Patterns, Pedagogy: A Corpus-Driven Approach to English Progressive Forms, Functions, Contexts, and Didactics* (Studies in Corpus Linguistics 18). Amsterdam: John Benjamins.

Römer, Ute. 2006a. Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 121–134.

Römer, Ute. 2006b. Looking at *looking*: Functions and contexts of progressives in spoken English and 'school' English. In Antoinette Renouf & Andrew Kehoe (eds.), *The Changing Face of Corpus Linguistics. Papers from the 24th International Conference on English Language Research on Computerized Corpora (ICAME 24)*, 231–242. Amsterdam: Rodopi.

Römer, Ute. 2007. Learner Language and the Norms in Native Corpora and EFL Teaching Materials: A Case Study of English Conditionals. In S. Volk-Birke & J. Lippert (eds.), *Proceedings of the Anglistentag 2006*, 355–363. Halle: Wissenschaftlicher Verlag Trier.

Römer, Ute. 2010. Using general and specialized corpora in English language teaching: Past, present and future. In Mari Carmen Campoy Cubillo, Begoña Bellés Fortuño

& Maria Lluïsa Gea-Valor (eds.), *Corpus-based approaches to English language teaching*, 18–38. London: Continuum.

Römer, Ute & Cynthia M. Berger. 2019. Observing the Emergence of Constructional Knowledge: Verb Patterns in German and Spanish Learners of English at Different Proficiency Levels. *Studies in Second Language Acquisition* 41(5). 1089–1110. https://doi.org/10.1017/S0272263119000202.

Rooy, Bertus van. 2006. The extension of the progressive aspect in Black South African English. *World Englishes* 25(1). 37–64. https://doi.org/10.1111/j.0083-2919.2006.00446.x.

Rose, David & James R. Martin. 2012. *Learning to Write, Reading to Learn: Genre, Knowledge and Pedagogy in the Sydney School*. London: Equinox.

Rühlemann, Christoph. 2008. A Register Approach to Teaching Conversation: Farewell to Standard English? *Applied Linguistics* 29(4). 672–693. https://doi.org/10.1093/applin/amn023.

Rundell, Michael. 2008. The corpus revolution revisited. *English Today* 24(1). 23–27. https://doi.org/10.1017/S0266078408000060.

Runte, Maren. 2015. *Lernerlexikographie und Wortschatzerwerb*. Berlin; Boston: Walter de Gruyter.

Rüschoff, Bernd & Dieter Wolff. 1999. *Fremdsprachenlernen in der Wissensgesellschaft: zum Einsatz der neuen Technologien in Schule und Unterricht*. Hueber Verlag.

Scales, Julie, Ann Wennerstrom, Dara Richard & Su Hui Wu. 2006. Language Learners' Perceptions of Accent. *TESOL Quarterly* 40(4). 715. https://doi.org/10.2307/40264305.

Schaer, Ursula. 2007. Source books rather than course books – Die Bildungsreform im Fremdsprachenunterricht und die neue Rolle für die Lehr-mittel. *Beiträge zur Lehrerbildung* 25(2). 255–267.

Schäfer, Werner. 2003. „Unterrichten ohne Lehrbuch? Einige unzeitgemäße Bemerkungen. *Praxis des neusprachlichen Unterrichts* 50(3). 305–311.

Schauer, Gila A. & Svenja Adolphs. 2006. Expressions of Gratitude in Corpus and DCT Data: Vocabulary, Formulaic Sequences, and Pedagogy. *System* 34(1). 119–134. https://doi.org/10.1016/j.system.2005.09.003.

Schegloff, Emanuel A. 1993. Reflections on Quantification in the Study of Conversation. *Research on Language & Social Interaction* 26(1). 99–128. https://doi.org/10.1207/s15327973rlsi2601_5.

Scheiwe, Lisa. in preparation. *Accents of English in ELT in Germany: A Corpus-Phonological Approach to Textbook Analysis*.

Schildhauer, Peter, Marion Schulte & Carolin Zehne. 2020. Global Englishes in the Classroom: From Theory to Practice. *PraxisForschungLehrer*innenBildung. Zeitschrift für Schul- und Professionsentwicklung.* 2(4). 26–40. https://doi.org/10.4119/pflb-3435.

Schlüter, Norbert. 2002. *Present perfect: eine korpuslinguistische Analyse des englischen Perfekts mit Vermittlungsvorschlägen für den Sprachunterricht* (Language in Performance 25). Tübingen: Narr.

Schmid, Hans-Jörg & Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems

and cognitive underpinnings. *Cognitive Linguistics* 24(3). https://doi.org/10.1515/cog-2013-0018.

Schmitt, Norbert & Ronald Carter. 2004. Formulaic sequences in action. In Norbent Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use*, 1–22. Amsterdam: John Benjamins.

Schönbrodt, Felix D. 2016. *p-hacker: Train your p-hacking skills!* https://shinyapps.org/apps/p-hacker/ (28 January, 2022).

Scott, Mike. 2011. *WordSmith Tools*. Stroud: Lexical Analysis Software.

Segermann, Krista. 2000. Eine neue Lehrwerk-Konzeption: Lehrbuch für Lehrer– Lernmaterialien für Schüler. *Praxis des neusprachlichen Unterrichts* 47(4). 339–348.

Seidlhofer, Barbara. 2001. Closing A Conceptual Gap: The Case For A Description Of English As A Lingua Franca. *International Journal of Applied Linguistics* 11(2). 133–158. https://doi.org/10.1111/1473-4192.00011.

Séré, Alain & Alain-Marie Bassy. 2010. *Le manuel scolaire à l'heure du numérique : Une « nouvelle donne » de la politique de ressources pour l'enseignement*. Inspection générale de l'éducation nationale. https://www.education.gouv.fr/le-manuel-scolaire-l-heure-du-numerique-1310 (6 January, 2022).

Shakir, Muhammad. 2020. *A corpus based comparison of variation in online registers of Pakistani English using MD analysis*. University of Münster PhD thesis. https://repositorium.uni-muenster.de/document/miami/1054a19c-abef-4d71-a324-c99ba7473da3/diss_shakir.pdf.

Siegel, A. 2014. What should we talk about? The authenticity of textbook topics. *ELT Journal* 68(4). 363–375. https://doi.org/10.1093/elt/ccu012.

Siepmann, Dirk. 2004. *Discourse Markers Across Languages: A Contrastive Study of Second-Level Discourse Markers in Native and Non-Native Text with Implications for General and Pedagogic Lexicography*. 1st edn. London: Routledge. https://doi.org/10.4324/9780203315262.

Siepmann, Dirk. 2007. Wortschatz und Grammatik: zusammenbringen, was zusammengehört. *Beiträge zur Fremdsprachenvermittlung* 46. 59–80.

Siepmann, Dirk. 2011. „Any Chance of a Bloody Drink Sometime This Century? Generische Strukturen einer Gesprächssituation analysieren. *Der fremdsprachliche Unterricht Englisch* 45(114). 22–26.

Siepmann, Dirk. 2014. Zur Repräsentation von Mehrwortausdrücken in deutschen Lehrwerken des Englischen. In Dirk Siepmann & Christoph Bürgel (eds.), *Sprachwissenschaft und Fremdsprachenunterricht: Spracherwerb und Sprachkompetenzen im Fokus*. Baltmannsweiler: Schneider Verlag Hohengehren.

Siepmann, Dirk. 2018a. *Volume Five: Prepositions* (A Grammar of Spoken and Written French). Leipzig: Amazon Distribution.

Siepmann, Dirk. 2018b. Zum Verhältnis von Fachwissenschaften, Fachdidaktik und Sprachpraxis in der universitären Lehrerausbildung: theoretische Überlegungen und Anregungen für eine kohärentere Praxis. In Bärbel Diehr (ed.), *Universitäre Englischlehrerbildung: Wege zu mehr Kohärenz im Studium und Korrespondenz mit der Praxis*, 103–121. Berlin: Peter Lang.

Siepmann, Dirk. 2019. *Band 3: Das Adjektiv* (Grammatik des gesprochenen und geschriebenen Französisch Band 3). Leipzig: Amazon Distribution.

Siepmann, Dirk & Christoph Bürgel. 2022. *Band 2: Das Nomen* (Grammatik des gesprochenen und geschriebenen Französisch). Leipzig: Amazon Distribution.

Siepmann, Dirk, John D. Gallagher, Mike Hannay & J. Lachlan Mackenzie. 2011. *Writing in English: A Guide for Advanced Learners*. 2nd, revised and extended edition edn. Tübingen Basel: Francke Verlag.

Sigley, Robert J. 1997. Text Categories and Where You Can Stick Them: A Crude Formality Index. *International Journal of Corpus Linguistics* 2(2). 199–237. https://doi.org/10.1075/ijcl.2.2.04sig.

Sinclair, John McH. 1983. Naturalness in Language. In Jan Aarts & Willem Meijs (eds.), *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research* (Costerus 45), 203–210. Leiden: Brill.

Sinclair, John McH. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John McH. 1992. The automatic analysis of corpora. In Jan Svartvik (ed.), *Directions in Corpus Linguistics*, 379–400. Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110867275.379.

Sinclair, John McH. (ed.). 2004. *How to Use Corpora in Language Teaching* (Studies in Corpus Linguistics). Amsterdam; Philadelphia: John Benjamins.

Sinclair, John McH., Gwyneth Fox, Stephen Bullon, Ramesh Krishnamurthy, Elisabeth Manning & John Todd (eds.). 1990. *Collins Cobuild English grammar: Helping learners with real English*. Glasgow: Harper Collins.

Sinclair, John McH. & Antoinette Renouf. 1988. A Lexical Syllabus for Language Learning. In Ronald Carter & Michael McCarthy (eds.), *Vocabulary and Language Teaching*, 140–158. Harlow: Longman.

Singapore Wala, Duriya Aziz. 2013. The Instructional Design of a Coursebook Is As It Is Because of What It Has To Do–An Application of Systemic Functional Theory. In Brian Tomlinson (ed.), *Developing Materials for Language Teaching*, 119–138. Second Edition. London ; New York: Bloomsbury.

Skehan, Peter. 2014. *Individual Differences in Second-Language Learning*. London: Taylor & Francis.

Smith, Christopher Arnold. 2020. *A Triangulated Accounting of Top Notch 2: Negotiating Ideologies in the Multimodal Discourse of an EFL Textbook in Korean University Classrooms*. Carleton University PhD thesis. https://curve.carleton.ca/00125d56-bc4c-4500-ab9a-d1fd4d8ccc02 (13 February, 2022).

Smith, Matthew. 2018. Dinner time or tea time? It depends on where you live. *YouGov*. https://yougov.co.uk/topics/politics/articles-reports/2018/05/22/dinner-time-or-tea-time-it-depends-where-you-live (29 September, 2020).

Smith, Nicholas. 2005. *A Corpus-based Investigation of Recent Change in the Use of the Progressive in British English*. Lancaster University PhD thesis. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.594241 (1 March, 2019).

Smith, Nicholas & Paul Rayson. 2007. Recent change and variation in the British English use of the progressive passive. *ICAME Journal* (31). 129–160.

Smitterberg, Erik. 2005. *The Progressive in 19th-Century English: A Process of Integration* (Language and Computers 54). Amsterdam: Rodopi.

Sommer, Roy. 2020. Lehrerbildung aus fachwissenschaftlicher Perspektive: Beispiel Anglistik. In Michaela Heer & Ulrich Heinen (eds.), *Die Stimmen der Fächer hören:*

Fachprofil und Bildungsanspruch in der Lehrerbildung, 307–324. Paderborn: Ferdinand Schöningh.

Spada, Nina & Patsy M. Lightbown. 1999. Instruction, First Language Influence, and Developmental Readiness in Second Language Acquisition. *The Modern Language Journal* 83(1). 1–22. https://doi.org/10.1111/0026-7902.00002.

Spoustová, Drahomíra, Jan Hajič, Jan Raab & Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron POS tagger. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*, 763–771. Athens, Greece: Association for Computational Linguistics. https://doi.org/10.3115/1609067.1609152. http://portal.acm.org/citation.cfm?doid=1609067.1609152 (12 August, 2021).

Spring, Ryan. 2018. Teaching Phrasal Verbs More Efficiently: Using Corpus Studies and Cognitive Linguistics to Create a Particle List. *Advances in Language and Literary Studies*. ERIC 9(5). 121–135.

Starkey, Louise, Miri Shonfeld, Sarah Prestridge & Mercè Gisbert Cervera. 2021. Special issue: Covid-19 and the role of technology and pedagogy on school education during a pandemic. *Technology, Pedagogy and Education*. Routledge 30(1). 1–5. https://doi.org/10.1080/1475939X.2021.1866838.

Stefanowitsch, Anatol. 2013. Variation and change in English path verbs and constructions: Usage patterns and conceptual structure. In Juliana Goschler & Anatol Stefanowitsch (eds.), *Human Cognitive Processing*, vol. 41, 223–244. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/hcp.41.10ste.

Stefanowitsch, Anatol. 2020. *Corpus Linguistics: A Guide to the Methodology* (Textbooks in Language Science 7). Berlin: Language Science Press.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Stein, Marcy, Carol Stuen, Douglas Carnine & Roger M Long. 2001. Textbook evaluation and adoption. *Reading & Writing Quarterly* 17(1). 5–23.

Stranks, Jeff. 2013. Materials for the Teaching of Grammar. In *Developing Materials for Language Teaching*, 337–350. Second edition. London; New York: Bloomsbury.

Sun, Zhuomin. 2010. Language Teaching Materials and Learner Motivation. *Journal of Language Teaching and Research* 1(6). 889–892. https://doi.org/10.4304/jltr.1.6.889-892.

Svartvik, Jan & Randolph Quirk. 1980. *A Corpus of English Conversation*. Lund: Studentlitteratur.

Swaffar, Janet K. 1985. Reading Authentic Texts in a Foreign Language: A Cognitive Model. *The Modern Language Journal* 69(1). 15–34.

Syndicat national de l'édition. 2021. Chiffres clés de l'édition. *Syndicat national de l'édition*. https://www.sne.fr/economie/chiffres-cles/ (2 January, 2022).

Tabachnick, Barbara G. & Linda S. Fidell. 2014. *Using Multivariate Statistics* (Always Learning). Pearson new international edition, sixth edition. Harlow: Pearson.

Tan, Melinda. 2003. Language corpora for language teachers. *Journal of Language and Learning* 1(2). 98–105.

Tateyama, Yumiko. 2019. Pragmatics in a Language Classroom. In Naoko Taguchi (ed.), *The Routledge Handbook of Second Language Acquisition and Pragmatics*, 400–413.

1st edn. London; New York, NY: Routledge. https://doi.org/10.4324/9781351164085-1.

Taylor, Charlotte. 2008. What is corpus linguistics? What the data says. *ICAME journal* 32. 179–200.

Tesch, Felicitas. 1990. *Die Indefinitpronomina Some Und Any Im Authentischen Englischen Sprachgebrauch Und in Lehrwerken: Eine Empirische Untersuchung*. Vol. 345. Tübingen: Gunter Narr Verlag.

Thompson, Paul, Susan Hunston, Akira Murakami & Dominik Vajn. 2017. Multi-Dimensional Analysis, text constellations, and interdisciplinary discourse. *International Journal of Corpus Linguistics* 22(2). 153–186. https://doi.org/10.1075/ijcl.22.2.01tho.

Thompson, Paul & Alison Sealey. 2007. Through children's eyes?: Corpus evidence of the features of children's literature. *International Journal of Corpus Linguistics* 12(1). 1–23.

Thornbury, Scott. 2000. Plenary: Deconstructing grammar. In Alan Pulverness (ed.), *IATEFL 2000: Dublin Conference Selections*, 59–67.

Thornbury, Scott. 2002. Training in instructional conversation. In Hugh Trappes-Lomax & Gibson Ferguson (eds.), *Language in Language Teacher Education* (Language Learning and Language Teaching 4), 95–106. Amsterdam: John Benjamins.

Timmis, Ivor. 2003. *Corpora, classroom and context: the place of spoken grammar in English language teaching*. University of Nottingham PhD thesis.

Timmis, Ivor. 2013. Corpora and Materials: Towards a Working Relationship. In Brian Tomlinson (ed.), *Developing Materials for Language Teaching*, 461–474. Second Edition. London; New York: Bloomsbury.

Timmis, Ivor. 2015. *Corpus Linguistics for Elt: Research and Practice* (Routledge Corpus Linguistics Guides). London; New York: Routledge. https://doi.org/10.4324/9781315715537.

Timmis, Ivor. 2016. Humanising coursebook dialogues. *Innovation in Language Learning and Teaching* 10(2). 144–153. https://doi.org/10.1080/17501229.2015.1090998.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work* (Studies in Corpus Linguistics v. 6). Amsterdam ; Philadelphia: J. Benjamins.

Tomasello, Michael. 2005. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, Mass.: Harvard University Press.

Tomlinson, Brian. 2001. Materials development. In David Nunan & Ronald Carter (eds.), *The Cambridge Guide to Teaching English to Speakers of Other Languages*, 66–71. Cambridge: Cambridge University Press.

Tomlinson, Brian (ed.). 2008. *English Language Learning Materials: A Critical Review*. London; New York: Continuum.

Tomlinson, Brian. 2012. Materials development for language learning and teaching. *Language Teaching* 45(02). 143–179. https://doi.org/10.1017/S0261444811000528.

Tomlinson, Brian (ed.). 2013a. *Applied Linguistics and Materials Development*. London; New York: Bloomsbury.

Tomlinson, Brian (ed.). 2013b. *Developing Materials for Language Teaching*. Second Edition. London; New York: Bloomsbury.

Tomlinson, Brian, Bao Dat, Hitomi Masuhara & Rani Rugdy. 2001. Survey review. EFL courses for adults. *ELT Journal* 55(1). 80–101. https://doi.org/10.1093/elt/55.1.80.

Tono, Yukio. 2004. Multiple Comparisons of IL, L1 and TL Corpora: The Case of L2 Acquisition of Verb Subcategorization Patterns by Japanese Learners of English. In Guy Aston, Silvia Bernardini & Dominic Stewart (eds.), *Studies in Corpus Linguistics*, vol. 17, 45–66. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.17.05ton.

Tony, Bex. 1999. *Standard English: The Widening Debate*. Hove: Psychology Press.

Torres-Martínez, Sergio. 2019. *Applied Cognitive Construction Grammar: A Cognitive Guide to the Teaching of Phrasal Verbs*. Medellín: Lulu.com.

Toutanova, Kristina, Dan Klein, Christopher D. Manning & Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 173–180. Association for Computational Linguistics.

Toutanova, Kristina & Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, vol. 13, 63–70. Hong Kong: Association for Computational Linguistics. https://doi.org/10.3115/1117794.1117802. http://portal.acm.org/citation.cfm?doid=1117794.1117802 (18 August, 2020).

Trabelsi, Soufiane, Brian Tomlinson & Hitomi Masuhara. 2010. Developing and trialling authentic materials for business English students at a Tunisian university. In *Research for Materials Development in Language Learning: Evidence For Best Practice*, 103–120.

Tracy-Ventura, Nicole & Magali Paquot. 2020. *The Routledge Handbook of Second Language Acquisition and Corpora*. London: Routledge.

Traugott, Elizabeth Closs & Graeme Trousdale. 2013. *Constructionalization and constructional changes* (Oxford Linguistics 6). First edition. Oxford: Oxford University Press.

Tsaroucha, Efthymia. 2018. *A Cognitive Linguistics Approach to English Phrasal Verbs*. Thessaloniki, Greece: Aristotle University of Thessaloniki Unpublished PhD Thesis. http://ikee.lib.auth.gr/record/299055.

Tyler, Andrea. 2012. *Cognitive Linguistics and Second Language Learning: Theoretical Basics and Experimental Evidence*. London: Routledge.

Tyler, Andrea E. & Lourdes Ortega. 2018. Usage-inspired L2 instruction: An emergent, researched pedagogy. In Andrea E. Tyler, Lourdes Ortega, Mariko Uno & Hae In Park (eds.), *Usage-Inspired L2 Instruction: Researched Pedagogy*, vol. 49, 3–26. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.49.01tyl.

Tyler, Andrea, Lourdes Ortega & Mariko Uno (eds.). 2018. *Usage-inspired L2 instruction: researched pedagogy* (Language Learning & Language Teaching 49). Amsterdam: John Benjamins.

UNESCO Institute for Statistics. 2012. *International Standard Classification of Education (ISCED) 2011.*

UNESCO Institute for Statistics/OECD/Eurostat. 2015. *ISCED 2011 Operational Manual : Guidelines for classifying national education programmes and related qualifications*. UNESCO Institute for Statistics/OECD/Eurostat. https://doi.org/10.15220/978-92-

9189-174-0-en. http://uis.unesco.org/sites/default/files/documents/isced-2011-operational-manual-guidelines-for-classifying-national-education-programmes-and-related-qualifications-2015-en_1.pdf (15 October, 2018).

Ur, Penny. 2011. Research, Theory, and Practice. In Eli Hinkel (ed.), *Handbook of research in second language teaching and learning. Vol. 2* (ESL & Applied Linguistics Professional Series), 507–522. New York, NY: Routledge.

Usó-Juan, Esther. 2008. A Pragmatic-Focused Evaluation of Requests and Their Modification Devices in Textbook Conversations. In Eva Alcón Soler (ed.), *Learning How to Request in an Instructed Language Learning Context*, 65–90. Bern: Peter Lang.

Usó-Juan, Esther & Alicia Martínez-Flor. 2010. The teaching of speech acts in second and foreign language instructional contexts. In *Pragmatics across Languages and Cultures*, 423–442. Berlin: Walter de Gruyter.

Valero Garcés, Carmen. 1998. Some pedagogical and practical implications of contrastive studies in ELT. *Revista española de lingüística aplicada* (13). 27–36.

Van Rossum, Guido & Fred L. Drake. 2009. Python 3 Reference Manual. Python Software Foundation. https://docs.python.org/3/reference/ (4 January, 2019).

Various users of neoprofs.org. 2016. [Anglais] La question des manuels : les utilisez-vous? Vous appuyez-vous sur celui de votre établissement? *www.neoprofs.org*. https://www.neoprofs.org/t103210-anglais-la-question-des-manuels-les-utilisez-vous-vous-appuyez-vous-sur-celui-de-votre-etablissement (6 January, 2022).

Varmış Kiliç, Zerhan & Binnur Genç İlter. 2015. The effect of authentic materials on 12th grade students' attitudes in EFL Classes. *ELT research journal* 4(1). 2–15.

Veirano Pinto, Marcia. 2019. Using Discriminate Function Analysis in Multi-Dimensional Analysis. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 217–230. Bloomsbury Academic. https://doi.org/10.5040/9781350023857.

Vellenga, Heidi. 2004. Learning Pragmatics from ESL & EFL Textbooks: How Likely? *TESL-EJ Teaching English as a Second or Foreign Language* 8(2). n. p.

Verspoor, Marjolijn. 2017. Complex dynamic systems theory and L2 pedagogy. In Lourdes Ortega & Han ZhaoHong (eds.), *Complexity Theory and Language Development: In Celebration of Diane Larsen-Freeman*, 143–62. Amsterdam: John Benjamins.

Verspoor, Marjolijn, Wander Lowie & Marijn Van Dijk. 2008. Variability in Second Language Development From a Dynamic Systems Perspective. *The Modern Language Journal* 92(2). 214–231. https://doi.org/10.1111/j.1540-4781.2008.00715.x.

Vielau, Axel. 2005. Lehrwerk, quo vadis? Einflüsse auf die Lehrwerkentwicklung. In Eva Burwitz-Melzer & Gert Solmecke (eds.), *Niemals Zu Früh und Selten Zu Spät: Fremdsprachenunterricht in Schule und Erwachsenenbildung. Festschrift für Jürgen Quetz*, 137–147. Berlin: Cornelsen.

Vine, Elaine W. 2013. Corpora and Coursebooks Compared: Category Ambiguous Words. In *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011)*, vol. 1, 463–478. Louvain-la-Neuve: Presses Universitaires de Louvain.

Virtanen, Tuija. 1997. The progressive in NS and NNS student compositions: evidence from the International Corpus of Learner English. In Magnus Ljung (ed.), *Corpus-based studies in English: Papers from the Seventeenth International conference on English*

Language Research on Computerized Corpora (ICAME 17). Stockholm, May 15–19, 1996, 299–309. Amsterdam: Rodopi.

Vogt, Karin. 2011. *Fremdsprachliche Kompetenzprofile: Entwicklung und Abgleichung von GeR-Deskriptoren für Fremdsprachenlernen mit einer beruflichen Anwendungsorientierung*. 1st edn. Tübingen: Gunter Narr Verlag.

Volansky, Vered, Noam Ordan & Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities* 98–118. https://doi.org/10.1093/llc/fqt031.

Volkmann, Laurenz. 2010. *Fachdidaktik Englisch: Kultur und Sprache*. Tübingen: Narr.

Vyatkina, Nina. 2020. Corpora as open educational resources for language teaching. *Foreign Language Annals* 53(2). 359–370. https://doi.org/10.1111/flan.12464.

Wain, Kenneth. 1992. Evaluating history and social studies textbooks. Whose criteria really does matter? In K. Peter Fritzche (ed.), *Schulbücher auf dem Prüfstand: Perspektiven der Schulbuchforschung und Schulbuchbeurteilung in Europa*, 23–31. Frankfurt: Moritz Diesterweg.

Wajnryb, Ruth. 1996. Death, taxes and jeopardy: Systematic omissions in EFL texts, or life was never meant to be an adjacency pair. Presented at the 9th Educational Conference, Sydney.

Wallis, Sean. 2020. *Statistics in Corpus Linguistics Research: A New Approach*. London: Routledge. https://doi.org/10.4324/9780429491696.

Weninger, Csilla & Tamas Kiss. 2013. Culture in English as a Foreign Language (EFL) Textbooks: A Semiotic Approach. *TESOL Quarterly* 47(4). 694–716. https://doi.org/10.1002/tesq.87.

Werfhorst, Herman van de, Emma Kessenich & Sara Geven. 2020. The digital divide in online education. Inequality in digital preparedness of students and schools before the start of the COVID-19 pandemic. pre-print published on OSF. https://osf.io/preprints/socarxiv/58d6p/ (1 February, 2022).

Werner, Valentin. 2021. Text-linguistic analysis of performed language: revisiting and re-modeling Koch and Oesterreicher. *Linguistics*. De Gruyter Mouton 59(3). 541–575. https://doi.org/10.1515/ling-2021-0036.

West, Michael. 1953. *A General Service List of English Words with Semantic Frequencies and a Supplementary Word-List for the Writing of Popular Science and Technology*. London: Longman.

Westergren Axelsson, Margareta & Angela Hahn. 2001. The use of the progressive in Swedish and German advanced learner English – a corpus-based study. *ICAME Journal* (25). 5–30.

Whitelaw, Casey & Shlomo Argamon. 2004. Systemic Functional Features in Stylistic Text Classification. *AAAI Technical Report* (7).

Widdowson, H. G. 1978. *Teaching Language as Communication*. Oxford: Oxford University Press.

Widdowson, H. G. 1984. *Explorations in Applied Linguistics 2*. Oxford: Oxford University Press.

Widdowson, H. G. 2003. *Defining Issues in English Language Teaching* (Oxford Applied Linguistics). Oxford: Oxford University Press.

Widdowson, Henry G. 1989. Knowledge of Language and Ability for Use. *Applied Linguistics* 10(2). 128–137. https://doi.org/10.1093/applin/10.2.128.

Wieling, Martijn, Josine Rawee & Gertjan van Noord. 2018. Reproducibility in Computational Linguistics: Are We Willing to Share? *Computational Linguistics* 44(4). 641–649. https://doi.org/10.1162/coli_a_00330.

Wiese, Heike, Katharina Mayr, Philipp Krämer, Patrick Seeger, Hans-Georg Müller & Verena Mezger. 2017. Changing teachers' attitudes towards linguistic diversity: effects of an anti-bias programme. *International Journal of Applied Linguistics* 27(1). 198–220. https://doi.org/10.1111/ijal.12121.

Wild, Kate, Adam Kilgarriff & David Tugwell. 2013. The Oxford Children's Corpus: Using a Children's Corpus in Lexicography. *International Journal of Lexicography* 26(2). 190–218. https://doi.org/10.1093/ijl/ecs017.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. Nature Publishing Group 3(1). 160018. https://doi.org/10.1038/sdata.2016.18.

Willis, Dave. 2003. *Rules, Patterns and Words: Grammar and Lexis in English Language Teaching*. Cambridge: Cambridge University Press.

Winter, Bodo. 2019. *Statistics for Linguists: An Introduction Using R*. New York: Routledge. https://doi.org/10.4324/9781315165547.

Winter, Tatjana & Elen Le Foll. forthcoming. Testing the Pedagogical Norm: Comparing *If*-conditionals in EFL Textbooks, Learner Writing and English Outside the Classroom. *International Journal of Learner Corpus Research*.

Wolff, Dieter. 1984. Lehrbuchtexte und Verstehensprozesse in einer zweiten Sprache. *Neusprachliche Mitteilungen aus Wissenschaft und Praxis* 37(1). 4–11.

Wolk, Christoph, Sandra Götz & Katja Jäschke. 2021. Possibilities and Drawbacks of Using an Online Application for Semi-automatic Corpus Analysis to Investigate Discourse Markers and Alternative Fluency Variables. *Corpus Pragmatics* 5. 7–36. https://doi.org/10.1007/s41701-019-00072-x.

Wong, S.C. 1983. Overproduction & Unidiomatic Usage in the Make Causatives of Chinese Speakers: A Cause for Flexibility in Interlanguage analysis. *Language Learning & Communication* 2(2). 151–165.

Wood, David. 2010. Lexical clusters in an EAP textbook corpus. In David Wood (ed.), *Perspectives on formulaic language: acquisition and communication*, 88–106. London ; New York: Continuum.

Wood, David & Randy Appel. 2014. Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes* 15. 1–13. https://doi.org/10.1016/j.jeap.2014.03.002.

Wulff, Stefanie, Nick C. Ellis, Ute Römer, Kathleen Bardovi-Harlig & Chelsea J. Leblanc. 2009. The Acquisition of Tense-Aspect: Converging Evidence From Corpora and Telicity Ratings. *The Modern Language Journal* 93(3). 354–369. https://doi.org/10.1111/j.1540-4781.2009.00895.x.

Wulff, Stefanie & Stefan Th. Gries. 2021. Exploring Individual Variation in Learner Corpus Research: Methodological Suggestions. In Bert Le Bruyn & Magali Paquot (eds.), *Learner Corpus Research Meets Second Language Acquisition* (Cambridge Applied Linguistics), 191–213. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108674577.010.

Wulff, Stefanie & Ute Römer. 2009. Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive. *Corpora* 4(2). 115–133. https://doi.org/10.3366/E1749503209000276.

Xiao, Richard. 2009. Theory-driven corpus research: Using corpora to inform aspect theory. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: an international handbook* (Handbücher Zur Sprach- Und Kommunikationswissenschaft = Handbooks of Linguistics and Communication Science), vol. 2, 987–1007. Berlin; New York: Walter de Gruyter.

Yoo, Isaiah WonHo. 2009. The English Definite Article: What ESL/EFL Grammars Say and What Corpus Findings Show. *Journal of English for Academic Purposes* 8(4). 267–278. https://doi.org/10.1016/j.jeap.2009.07.004.

Zarifi, Abdolvahed & Jayakaran Mukundan. 2012. Phrasal Verbs in Malaysian ESL Textbooks. *English Language Teaching* 5(5). 9–18. https://doi.org/10.5539/elt.v5n5p9.

Zipf, George Kingsley. 1935. *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: MIT Press.

Zuppardo, Maria Carolina. 2013. A linguagem da aviação: um estudo de manuais aeronáuticos baseado na Análise Multidimensional [Aviation Language: A Study of Aeronautical Handbooks Based on Multi-Dimensional Analysis]. *Revista Virtual de Estudos da Linguagem* 11. 6–25.

Appendices

The **Online Appendix** can be found on:
https://elenlefoll.github.io/TextbookEnglish.

It is linked to the project's GitHub repository:
https://github.com/elenlefoll/TextbookEnglish

Category	Feature	Code/ Tag	Examples	Operationalisation	Normalisation unit	As coded by
Features for which there are no tags in the tagged texts						
General text properties	Total number of words	**Words**	*It's a shame that you'd have to pay to get that quality.* (= 14)	The number of tokens as tokenised by the Stanford Tagger, but excluding punctuation marks, brackets, symbols, genitive 's (POS), and filled pauses and interjections (FPUH). Contractions are treated as separate words, i.e.,it's is tokenised as it and 's. Note that this variable is only used to normalise the frequencies of other linguistic features.	NA	Le Foll
General text properties	Average word length	**AWL**	*It's a shame that you'd have to pay to get that quality.* (42/12 = 3.50)	Total number of characters in a text divided by the number of words in that same text (as operationalised in the Words variable above, hence excluding filled pauses and interjections, cf. FPUH).	Words	Le Foll
General text properties	Lexical diversity	**TTR**	***It's a shame that you'd have to pay*** *to* ***get*** *that* ***quality*** *.* (12/14 = 0.85)	Following Biber (1988), this feature is a type-token ratio measured on the basis of, by default, the first 400 words of each text only. It is thus the number of unique word forms within the first 400 words of each text divided by 400. This number of words can be adjusted in the command used to run the script (see instructions at the top of the MFTE script).	Words (by default first 400)	Le Foll
General text properties	Lexical density	**LDE**	*It's a* ***shame*** *that you'd have to* ***pay*** *to get that* ***quality*** *.* (3/14 = 0.21)	For this feature, tokens which are not on the list of the 352 function words from the {qdapDictionaries} R package, nor individual letters, or any of the fillers listed in FPUH are identified as content words. Lexical density is calculated as the ratio of these content words to the total number of words in a text.	Words	Le Foll
General text properties	Finite verbs	**FV**	*He discovered that the method involved imbiding copious amounts of tea. Ants can survive by joining together to morph into living rafts. Always wanted to experience the winter wonderland that Queen Elsa created?*	This feature is not directly listed in the MFTE output tables; however, it is used as a normalisation basis for many other linguistics features (see Normalisation column). It is calculated by tallying the number of occurrences of the following features: VPRT, VBD, VIMP, MDCA, MDCO, MDMM, MDNE, MDWO and MDWS.	NA	Le Foll
Features for which there are tags in tagged version of the texts processed by the MFTE						
Adjectives	Attributive adjectives	**JJAT**	*I've got a* ***fantastic*** *idea! I didn't sleep at all* ***last*** *night.* ***Cheap*** *,* ***quick*** and ***easy*** *fix!*	Whereas the Biber Tagger and the MAT first identify predicative adjectives and then consider all remaining J.* tags from the Stanford Tagger to be attributive adjectives, the MFTE proceeds the other way around because it is considerably easier to reliably identify attributive adjectives than it is predicative adjectives. Thus, all adjectives (J.*, as tagged by the Stanford Tagger) followed by another adjective, a noun or a cardinal number, or preceded by a determiner are tagged as attributive adjectives. Once these first attributive adjectives have been identified, an additional loop is run to capture any additional attributive adjectives found in lists of attributive adjectives.	Nouns	Le Foll

Adjectives	Predicative adjectives	**JJPR**	*That's **right**. One of the main advantages of being **famous** ... It must be absolutely **wonderful**.*	Once attributive adjectives have been identified (see JJAT) and tagged as JJAT, all remaining JJ, JJS and JJR tags are overwritten as JJPR. In addition, *ok* and *okay* in the construction *BE ok(ay)* are also tagged as JJPR. These words are otherwise identified as foreign words (FW) by the Stanford Tagger.	Finite verbs	Le Foll
Adverbials	Frequency references	**FREQ**	*We should **always** wear a mask. But he had found his voice **again.***	Assigned to all occurrences of the frequency adverbs listed in the COBUILD (Sinclair et al. 1900: 270): *usually, always, mainly, often, generally, normally, traditionally, again, constantly, continually, frequently, ever, never, infrequently, intermittently, occasionally, often, periodically, rarely, regularly, repeatedly, seldom, sometimes* and *sporadically.*	Finite verbs	Le Foll
Adverbials	Place references	**PLACE**	*It's not **far** to go. I'll get it from **upstairs**. It's **downhill** all the way. It's **there** not here.*	Biber's (1988: 224) list of place adverbials was taken from Quirk et al. (1985:514ff) but inexplicably excludes many from this list. Those that do not fulfil other major functions were therefore added: downwind, eastward(s), westward(s), northward(s), southward(s), upwards, downwards, elsewhere, everywhere, here, offshore, nowhere, somewhere, thereabout(s) and there (but occurrences of there tagged as existential there (EX) by the Stanford Taggers were ignored). Only occurrences of far which have not previously identified as TIME references (e.g., so far, thus far) or emphatics (e.g., far better, far more) are tagged as PLACE references.	Finite verbs	Le Foll, adapted from Biber (1988)
Adverbials	Time references	**TIME**	*It will **soon** be possible. **Now** is the time. I haven't come across any issues **yet**.*	All occurrences of *afterwards, again, earlier, early, eventually, forever, formerly, immediately, initially, instantly, late, lately, later, momentarily, now, nowadays, once, originally, presently, previously, recently, shortly, simultaneously, subsequently, today, to-day, tomorrow, to-morrow, tonight, to-night, yesterday.* Following Nini (2014: 18), the word *soon* was not tagged as a time adverbial when followed by the word *as*. *Ago, already, beforehand, prior to,* and *far* (the latter only when proceeded by *so* or *thus* and not followed by an adjective or adverb), and *am* and *pm* as adverbs were added to the list, as well as *yet* tokens that have not previously been identified as concessives (CONC).	Finite verbs	Le Foll, adapted from Nini (2014)
Adverbials	Other adverbs	**RB**	***Unfortunately** that's the case. **Exactly** two weeks. He could **so easily** but he knows **better**. He's **still** gonna come **back**.*	Corresponds to all the tokens tagged as RB, RBS, RBR or WRB by the Stanford Tagger apart from those identified as adverbs of frequency (FREQ), place (PLACE) or time (TIME), amplifiers (AMP), emphatics (EMPH), hedges (HDG) and downtoners (DWNT).	Words	Le Foll
Determinatives	*s* -genitives	**POS**	*the world **'s** two most populous country, my parents **'** house*	As identified by the Stanford Tagger: the possessive endings on nouns ending in 's or '. Note that these tokens are not counted as Word in the computation of the lexical diversity (TTR) and average word length variables (AWL) features.	Nouns	Le Foll
Determinatives	Determiners	**DT**	*Is that **a** new top? **The** first line has to be interesting. Are they **both** Spice Girls? On **either** side of the page. To **another** room. They're five pounds **each**.*	As tagged by the Stanford Tagger (DT) (Santorini 1990: 2), with the exception of *that, this, these* and *those* which are counted as demonstratives (DEMO). Note that this Stanford Tagger category also includes pronouns such as *another* in *Shall I choose <u>another</u> ?*	Nouns	Le Foll

Determinatives	Quantifiers	**QUAN**	***Such*** *a good time in like* ***half*** *an hour. She's got* ***all*** *these great ideas. It happens* ***each*** *and* ***every*** *time.*	All occurrences of pre-determiners as tagged by the Stanford Tagger, which includes the following "determiner-like elements when they precede an article or possessive pronoun" (Santorini 1990: 4): *nary, quite, rather* and *such* (e.g., quite a mess, rather a nuisance, many a moon), as well as all instances of all (unless immediately followed by *right* , cf. DMA), *any, a bit, both, each, every, few, half, many, much, several, some, lots, a lot (of), load(s) of, heaps of, wee, less* and *more* (as adjectives only).	Nouns	Le Foll
Determinatives	Numbers	**CD**	*That's her number* ***one*** *secret.* ***Two*** *eyes glowed just above the surface. It happened on* ***7*** *February,* ***2019*** *.*	All cardinal numbers as identified by the Stanford Tagger. This includes dates written in numbers, e.g., *1994.* In addition, numbers listed as list markers (LS) by the Stanford are overwritten as CD and strings of the type *\b[0-9]+th_ | \b[0-9]+nd_ | \b[0-9]+rd_* are also tagged as numbers (CD).	Words	Le Foll
Determinatives	Demonstratives	**DEMO**	*What are you doing* ***this*** *weekend? I love* ***that*** *film. Whoever did* ***that*** *should admit it.*	Assigned to all occurrences of *that, this, these* and *those* identified by the Stanford Tagger as determiners (DT).	Words	Le Foll
Discourse organisation	Elaborating conjunctions	**ELAB**	***Similarly*** *, you may,* ***for example*** *, write bullet points* ***insomuch as*** *it helps you to focus your ideas.*	*Assigned to such that (not followed by a determiner), such as, inasmuch as, insofar as, insomuch as, in that, to the extent that, in particular, in conclusion, in sum, in summary, to summarise, to summarize, for example, for instance, in fact, in brief, in any event, in any case, in other words, e(.)g(.), in summary, viz(.), cf(.), i.e., namely, etc(.), likewise, namely, as well as similarly and accordingly when followed by a comma.*	Finite verbs	Le Foll
Discourse organisation	Coordinating conjunctions	**CC**	***Instead*** *of listening to us, he* ***also*** *told John* ***and*** *Jill* ***but*** *at least his parents don't know yet.*	*This category takes the coordinating conjunctions (CC) tagged by the Stanford Tagger as its basis which include and, but, nor, or, yet, "as well as the mathematical operators plus, minus, less, times (in the sense of 'multiplied by') and over (in the sense of 'divided by'), when they are spelled out" (Santorini 1990: 2). However, conjunctions already captured by other variables are excluded from this count: yet is assigned to concessive (CONC). In addition, the following (multi-word) conjunctions are also included in this category: also, besides, moreover, further (when tagged as an adverb), furthermore, in addition, additionally, as well (as) (except when preceded by least), however (provided it is preceded or followed by a punctuation mark), ibid, on the one hand, on the other hand, instead, besides, conversely, by/in contrast, on the contrary, in/by comparison, whereas, whereby, whilst.*	Finite verbs	Le Foll
Discourse organisation	Causal conjunctions	**CUZ**	*He was scared* ***because*** *of the costume. Yeah* ***coz*** *he hated it.*	Assigned to all occurrences of *because* , *'cause, cos, cuz* and *coz* . The latter four were not included in Biber's (1988) original variable. According to Biber (1988: 236) *because* "is the only subordinator to function unambiguously as a causative adverbial". Whilst it is true that many subordinators, e.g., *as, for,* and *since* , can fulfil a range of functions, including causative, and were therefore not included in this category, the following adverbs and multi-word conjunctions were added since they mostly fulfil a causative function: *as a result, on account of, for that/this purpose, thanks to, to that/this end, consequently, in consequence, hence, so that, therefore, thus* .	Finite verbs	Le Foll, adapted from Biber (1988)

Discourse organisation	Concessive conjunctions	**CONC**	*Even **though** the antigens are normally hidden…*	Assigned to all occurrences of *although, though, tho, despite, except that, in spite of, albeit, granted that, nevertheless, nonetheless, notwithstanding, whereas, no matter +* WH-word*, (ir)regardless of* , and *granted.* Also assigned to *still* and *yet* when preceded by any punctuation mark or followed by a comma. Multi-word units are only counted as one occurrence of CONC.	Finite verbs	Le Foll
Discourse organisation	Conditional conjunctions	**COND**	***If** I were you… Even **if** the treatment works…*	Assigned to all occurrences of *if, as long as, unless, lest, in that case, otherwise, whether.*	Finite verbs	Le Foll
Discourse organisation	Discourse/pragmatic markers	**DMA**	***Well no** they didn't say **actually** . **Okay I guess** we'll see how things go **right** ?*	Assigned to "interactional signals and discourse markers" (as listed in Stenström 1994: 59 and cited in Aijmer 2002: 2): *actually, all right, anyway, God, goodness, gosh, OK, okay, right* (if tagged as an interjection by the Stanford Tagger)*, well* (only if identified by the Stanford Tagger as an adverb or adjective and not if preceded by *as, how, very, really, quite,* a verb, an adjective or an adverb)*, yes, yeah, yep, sure* (unless it is preceded by the verb *MAKE, for, not* or *you*). Verbal phrases such as *you know* and *I mean* were excluded from this variable since literal occurrences could not be automatically disambiguated occurrences as discourse markers. A number of markers from Stenström's list are also not assigned this tag because they are captured by other variables: *now* (TIME), *please* (POLITE), *really* (EMPH), *quite* and *sort of* (HDG). The following items were added: *lol, IMO, omg, wtf, nope, mind you, of course, whatever* and *damn* (unless tagged as a verb, or followed by an adjective; in the latter case it is an emphatic, cf. EMPH).	Words	Le Foll
Discourse organisation	Filled pauses and interjections	**FPUH**	***Oh** nooooooo, Tiger's furious! **Wow** ! **Hey** Tom! **Er** I don't know. **Hmm** .*	Assigned to all occurrences of ah+, aw+, oh+, eh+, er+, erm+, mm+, ow+, um+, huh+, uhu+, uhuh, mhm+, hm+ (but not HM), oo+ps woo+ps, hi, hey, and interjections identified by the Stanford Tagger and not assigned to another category. The plus sign (+) signifies that that the preceding letter can appear multiple times, i.e.,ahh and errrr are also assigned this tag.	Words	Le Foll
Discourse organisation	*Like*	**LIKE**	*Sounds **like** me. And just **like** his father. And he was **like** this isn't true. I wasn't gonna **like** do it.*	Occurrences of like tagged as a preposition (IN) or adjective (JJ) by the Stanford Tagger are assigned this tag because, in spoken English, like typically fulfils a range of different functions, e.g.,fillers and softeners, and attempts to disambiguate like as a preposition or conjunct proved too error-prone. This category excludes occurrences of like identified as the quotative BE + like (QLIKE) if the QLIKE feature is included (which, by default, it is not, cf. tagger evaluation).	Words	Le Foll
Discourse organisation	*So*	**SO**	*She had spent **so** many summers there. **So** there you go.*	Occurrences of *so* tagged as IN by the Stanford Tagger and not previously identified as either an emphatic (*so +* J.**/much/many/little* ; EMPH) or an adverbial subordinator (*so that* + NN.*/J.*; OSUB) are assigned this tag.	Words	Le Foll
Discourse organisation	Direct WH-questions	**WHQU**	***What**'s happening? **Why** don't we call the game off? **How** ? And **who** is Dinah, if I might venture to ask the question?*	Assigned to *what, where, when, how, why, who, whom, whose* and *which* followed by a question mark within 15 tokens.	Finite verbs	Le Foll

Discourse organisation	Question tags	**QUTAG**	*Do they **?** Were you **?** It's just it's repetitive, isn't it **?***	Assigned to question marks preceded by (1) *innit, init;* (2) a modal verb (MD) or *did* or *had,* and a personal pronoun (P.+); (3) a modal verb or *did* or *had,* a negation (XX0), and a personal pronoun; (4) *is, does, was* or *has* , followed by *it, she* or *he;* (5) *is, does, was* or *has* , followed by a negation, and *it, she* or *he* ; (6) *do, were, are* or *have* , followed by *you, we* or *they* ; (7) *do, were, are* or *have* , followed by a negation, and *you, we* or *they* . In addition, the above patterns are not considered question tags if a question word occurs within six words to the left of the question mark; consequently, *Why did you do it?* is not assigned this tag but rather WHQU.	Finite verbs	Le Foll
Discourse organisation	Yes/no questions	**YNQU**	***Have** you thought about giving up? **May** I take a seat? **Do** you mind?*	Assigned to any form of the verbs BE, HAVE, DO or a modal verb (MD) followed by a personal pronoun (P.+), a noun (NN.*), a negation (XX0) or determiner (DT) and then a question mark within three to 15 tokens, as long as no WH-question (WHQU) or yes/no question tag (YNQU) is present one or two tokens before the auxiliary verb. Note that this variable should not overlap with question tags (QUTAG).	Finite verbs	Le Foll
Discourse organisation	*that* relative clauses	**THRC**	*You must be very clever to find a use for something **that** costs nothing. I'll just run a cable **that** goes from here to there.*	Assigned to that identified as introducing a relative clause by the Stanford Tagger (WDT), unless it is immediately followed by a punctuation mark. Any remaining that_WDT tokens are typically mistagged demonstratives and are thus assigned to the DEMO category, e.g., *I don't think that's a problem that is.*	Finite verbs	Le Foll
Discourse organisation	*that* subordinate clauses (other than relatives)	**THSC**	*Did you know **that** the calendar we use today was started by Julius Caesar? She resented being told constantly **that** she was ignorant and stupid.*	Assigned to that tokens which have been tagged as IN by the Stanford Tagger and are not immediately followed by a punctuation mark. Remaining that_IN tokens are assigned to the demonstrative category (DEMO): these are end-of-sentences/utterances tokens which are typically misidentified by the Stanford Tagger, e.g., *Who was that?*	Finite verbs	Le Foll
Discourse organisation	Subordinator *that* omission	**THATD**	*I mean **[THATD]** you'll do everything. I thought **[THATD]** he just meant our side. You don't think **[THATD]** he's a drug dealer? I know **[THATD]** that's not his thing.*	The THATD tag is assigned to the following patterns: (1) a public, private or suasive verb followed by a demonstrative pronoun (DEMO) or I, we, he, she, it, they and then a verb (V.* or MD); (2) a public, private or suasive verb followed by I, we, he, she, it, they or a noun (N.*), and then by a verb (V.* or MD); (3) a public, private or suasive verb followed by an adjective (J.*), an adverb (RB), a determiner (DT, QUAN, CD) or a possessive pronoun (PRPS), and then a noun (N.*), and then a verb (V.* or MD), with the possibility of an intervening adjective (J.*) between the noun and its preceding word. This tag corresponds to Biber's (1988: 244) category but its operationalisation has been improved to avoid the algorithm erroneously tagging constructions such as *Why would I know that?* and *He didn't hear me thank God.*	Finite verbs	Le Foll, adapted from Biber (1988)
Discourse organisation	WH subordinate clauses	**WHSC**	*I'm thinking of someone **who** is not here today. Do you know **whether** the banks are open?*	Assigned when the words *what, where, when, how, whether, why, whoever, whomever, whichever, wherever* and *whenever* have not been previously identified as part of a WH question (WHQU). Though many attempts were made, it proved impossible to reliably disambiguate between relative and other subordinate WH-clauses, which is why they are pooled together in this category.	Finite verbs	Le Foll

Lexis	Total nouns (including proper nouns)	**NN**	*a **cut,** my **coat,** the **findings,** cruelty **, comprehension,** on **Monday** 6 **Aug,** the **U.S.** , on the **High Street***	Assigned to all singular (NN) and plural nouns (NNS) identified by the Stanford Tagger including proper nouns (NNP and NNPS). This variable differs from the Biber Tagger in that it includes nominalisations.	Words	Le Foll
Lexis	Noun compounds	**NCOMP**	*Surely this stone must be the last one to cover the dungeon **entrance** ! Experts say that the rare winter **phenomenon** is a natural occurrence.*	Assigned when two or more nouns follow each other without any intervening punctuation. The algorithm allows for the first noun to be a proper noun but not the second thus allowing for *Monday afternoon* and *Hollywood stars* but not *Barack Obama* and *Los Angeles* . It is also restricted to nouns with a minimum of two letters to avoid OCR errors (dots and images identified as individual letters and which are usually tagged as nouns by the Stanford Tagger) producing too many erroneous NCOMP's. Note that this feature works best with fully punctuated texts (see per-register recall and precision rates in the tagger documentation).	Nouns	Le Foll
Lexis	Emoji and emoticons	**EMO**	😍 🥰 🌈 :-) :DD XD <3 :/	Assigned to all emojis as of December 2018 (cf. https://unicode.org/emoji/charts/full-emoji-list.html) and to a range of emoticons, in particular three-character emoticons such as *:-)* . The source code also includes three lines which are by default commented out but can be uncommented for texts where short emoticons are expected. It is not recommended to use these lines for general English because they lead to a sharp decrease in precision: many of the shorter emoticons, e.g., *:(:D :3* , are too easy to confuse with poorly scanned texts that are missing spaces, or with the punctuation styles of specific academic journals.	Words	Le Foll
Lexis	Hashtags	**HST**	***#phdlife #Buy1Get1Free***	Assigned to any string starting with a hashtag followed by at least three letters, digits or underscores.	Words	Le Foll
Lexis	URL and e-mail addresses	**URL**	***www.faz.net https://twitter.com elefoll@uos.de***	Assigned to all strings resembling a URL or an e-mail address (without claiming to only include valid URLs or e-mail addresses since this is not the aim). Regex for this feature was inspired by: https://mathiasbynens.be/demo/url-regex	Words	Le Foll
Negation	Negation	**XX0**	*Why do **n't** you believe me? There is **no** way that's happening any time soon. **Nor** am I.*	Biber's (1988) analytic and synthetic negation features were merged into one negation variable since the latter is too infrequent to be of use in the context of this study. This unique negation tag is assigned to the tokens *not_RB, n't* _RB, all occurrences of the words *nor* and *neither* , and *no* when followed by an adjective (J.*) or noun (NN.*).	Finite verbs	Le Foll
Prepositions	Prepositions	**IN**	*The Great Wall **of** China is the longest wall **in** the world. There are towers **along** the wall. I prefer **to** go **to** an art gallery. The objects **on** display are **from** all **over** the world.*	All items tagged as IN by the Stanford Tagger other than those assigned to CAUS, CONC, COND, OSUB, SO and LIKE.	Words	Le Foll

Pronouns	Reference to the speaker/writer	**FPP1S**	***I*** *don't know. It isn't* ***my*** *problem.*	All occurrences of *me, myself* and *mine* and *I* if tagged by the Stanford Tagger as a pronoun, a list symbol (LS) or a foreign word (FW).	Finite verbs	Le Foll
Pronouns	Reference to the speaker/writer and other(s)	**FPP1P**	***We*** *were told to deal with it* ***ourselves*** *.*	All occurrences of us, we, our, ourselves and ours, as well as the contracted form of us (e.g., in let's). All these terms are case insensitive but an exception for US was added as this usually refers to the United States of America.	Finite verbs	Le Foll
Pronouns	Reference to the addressee	**SPP2**	*If* ***your*** *model was good enough,* ***you*** *'d be able to work it out.*	Following Biber (1988), all occurrences of *you, your, yourself, yourselves* . Following Nini (2014: 18), also includes *thy, thee* and *thyself* . In addition, the forms *ur, ye, y'all, ya* , *thine* and the nominal possessive pronoun *yours* were also added.	Finite verbs	Le Foll, adapted from Nini (2014)
Pronouns	*it* pronoun reference	**PIT**	***It*** *fell and broke. I implemented* ***it*** *.* ***Its*** *impact has not yet been researched.*	All occurrences of the pronoun *it* . An exception was added for the all capital form *IT* which most frequently refers to *Information Technology* . Following Nini (2014: 18), also includes all occurrences of *itself* and *its* .	Finite verbs	Le Foll, adapted from Nini (2014)
Pronouns	*One* as a personal pronoun	**PRP**	***One*** *would hardly suppose that your eye was as steady as ever.*	This tag consists of the remaining personal pronouns not yet tagged as either first (FPP1S and FPP1P), second (SPP2) or third (TPP3) person pronouns. In practice, this should only leave *one* .	Finite verbs	Le Foll
Pronouns	Reference to one non-interactant	**TPP3S**	***He*** *is beginning to form* ***his*** *own opinions.* ***She*** *does tend to keep to* ***herself*** *.*	Following Biber (1988), all occurrences of *she, he, her, him, his, himself, herself* and *themself* . Note that the singular *they* form can only be accounted for with the possessive pronoun: *themself.*	Finite verbs	Le Foll
Pronouns	Reference to more than one non-interactant	**TPP3P**	*The text allows readers to grapple with* ***their*** *own conclusions. I wouldn't trust* ***them*** *.*	All occurrences of *they, them, themselves, theirs* and *em* when tagged by the Stanford Tagger as a pronoun.	Finite verbs	Le Foll
Pronouns	Quantifying pronouns	**QUPR**	*said Alice aloud, addressing* ***nobody*** *in particular.*	All occurrences of *anybody, anyone, anything, each other, everybody, everyone, everything, nobody, none, no one, nothing, somebody, someone* and *something.*	Finite verbs	Nini (2014)
Stance-taking devices	Politeness markers	**POLITE**	*Can you open the window,* ***please*** *? Would you* ***mind*** *giving me a hand? I was* ***wondering*** *whether you could help.*	Assigned to all occurrences of *thanks, thank you, cheers, ta* (unless it is preceded by *got* to avoid the confusion with *gotta*), *please, sorry, apology, apologies,* all forms of the verbs *excuse, I/we wonder, I/we + BE + wondering,* and the multi-word units *you mind* and *don't mind* . No exception was made for *please* as a verb because the Stanford Tagger frequently misidentifies *please* as a verb, e.g., *I was like please_VPRT just please_VB just get there.*	Words	Le Foll

Stance-taking devices	Amplifiers	**AMP**	*I am **very** tired. They were both **thoroughly** frightened.*	Assigned to the amplifiers from Biber's (1988) list: *absolutely, altogether, completely, enormously, entirely, extremely, fully, greatly, highly, intensely, perfectly, strongly, thoroughly, totally, utterly, very. Especially* was added.	Words	Le Foll, adapted from Biber (1988)
Stance-taking devices	Downtoners	**DWNT**	*These tickets were **only** 45 pounds. It's **almost** time to go.*	Assigned to all occurrences of *almost, barely, hardly, merely, mildly, nearly, only, partially, partly, practically, scarcely, slightly, somewhat* . In Biber (1988) *almost* is listed as both a hedge and a downtoner. Following Nini (2014), it is only considered a downtoner here.	Words	Nini (2014)
Stance-taking devices	Emphatics	**EMPH**	*I **do** wish I hadn't drunk quite **so** much. Oh **really** ? I **just** can't get my head around it.*	Following Biber (1988), assigned to all occurrences of *just, really, most, more, real* + ADJ, *so* + ADJ, *for sure, such a* . The algorithm was improved by adding *so + much/little/many, such a/* an (whilst excluding *such a/an* if proceeded by *of*), and ensuring that only *DO* + verb in base form (VB) are tagged. *Least* and *far* + J.*/RB were added (the latter only when not proceeded by *so* or *thus*). To account for recent language change (Aijmer 2018), *bloody, dead* + ADJ, *fucking* and *super* were also added. Multi-word units are counted as one EMPH tag but several Words.	Words	Le Foll, adapted from Biber (1988)
Stance-taking devices	Hedges	**HDG**	*There seemed to be no **sort** of chance of getting out. I wish that **kind** of thing never happened. She's **maybe** gonna do it.*	Following Biber (1988: 240) assigned to all occurrences of *maybe, at about, something like,* and *more or less,* as well as *sort of* and *kind of* as long as they are not preceded by a determiner (DT), quantifier (QUAN), cardinal number (CD), adjective (J.*), possessive pronoun (PRPS) or WH word. The condition that *kind* must have been tagged as a noun (NN) by the Stanford Tagger was added to exclude phrases such as *it's very kind of you* . *Kinda* and *sorta* was added as colloquial alternatives to *kind of* and *sort of* and the adverbs *apparently* , *conceivably, perhaps, possibly, presumably, probably, roughly* and *somewhat* were also added to the list.	Words	Le Foll, adapted from Biber (1988)
Stative forms	Existential *there*	**EX**	***There** are students. And **there** is now a scholarship scheme.*	As tagged by the Stanford Tagger: "Existential *there* is the unstressed *there* that triggers inversion of the inflected verb and the logical subject of a sentence" (p. 3).	Finite verbs	Le Foll
Stative forms	Be as main verb	**BEMA**	*It was nice to just **be** at home. She'**s** irreplaceable. It'**s** best I think. How **was** your mum on Sunday? It'**s** not long.*	Following Biber (1988), this tag is assigned to the all forms of the verb be when followed by a determiner (DT), a possessive pronoun (PRPS) a preposition (IN), or an adjective (JJ). In addition, Nini (2014: 20) improved the Biber Tagger "by taking into account that adverbs or negations can appear between the verb BE and the rest of the pattern. Furthermore, the algorithm was slightly modified and improved: (a) the problem of a double-coding of any Existential there followed by a form of BE as a BEMA was solved by imposing the condition that there should not appear immediately before or two before the pattern; (b) the cardinal numbers (CD) tag and the personal pronoun (PRP) tag were added to the list of items that can follow the form of BE." This latter improvement by Nini, however, resulted in tag questions also being assigned to BEMA. The present algorithm therefore further excludes any occurrences of BE found one or two to the left of a question tag (QUTAG), as well as BE occurrences one or two to the left of a present participle form tagged as PROG or past participle form tagged as PASS.	Finite verbs	Le Foll, adapted from Nini (2014)

Syntax	Split auxiliaries and infinitives	**SPLIT**	*I would actually **drive** . You can just so **tell** . I can't ever **imagine** arguing with Jill.*	This category merges Biber's (1988) split auxiliaries and split infinitive categories and follows Nini's (2014: 30) operationalisations. Hence, this tag is assigned every time the infinitive marker to (TO) is followed by one or two adverbs and a verb base form, and every time an auxiliary (any modal verb MD, or any form of DOAUX, or any form of BE, or any form of HAVE) is followed by one or two adverbs and a verb form. Nini's algorithm was improved to ensure that negated split auxiliaries would also be identified, e.g., They have not yet developed cancer.	Finite verbs	Le Foll, adapted from Nini (2014)
Syntax	Stranded prepositions	**STPR**	*We've got more than can be accounted **for** . Open the door and let them **in** . Where is it **from** ? It's not the sort of music we're **into** .*	As in Biber (1988), assigned to the prepositions *against, amid, amidst, among, amongst, at, between, by, despite, during, except, for, from, in, into, minus, of, off, on, onto, opposite, out, per, plus, pro, than, through, throughout, thru, toward, towards, upon, versus, via, with, within* and *without* followed by any punctuation mark. Following Nini (2014: 30), *besides* was removed from Biber's original list since it also frequently serves as a conjunct and, in this function, is usually followed by a punctuation mark. Note that Nini's (2014:30) operationalisation tagged all occurrences of these word forms as prepositions regardless of how they were tagged by the Stanford Tagger. Here, it was decided to improve accuracy by restricting the query to tokens tagged as IN by the Stanford Tagger (thus excluding many RB and RP tokens, e.g., *Don't take it away ! Tie her up ! He roared out : "Come away !"*).	Finite verbs	Le Foll, adapted from Nini (2014)
Verb features	Verbal contractions	**CONT**	*I do **n't** know. It is **n't my** problem. You **'ll** have to deal with it.*	Following (Nini 2014: 29), all occurrences of an apostrophe followed by a word identified as a verb (V.*, MD) by the Stanford Tagger and all occurrences of the token *n't* _XX0.	Finite verbs	Nini (2014)
Verb features	Particles	**RP**	*I'll look it **up.** It's coming **down** . When will you come **over** ? Some of the birds hurried **off** at once.*	As tagged by the Stanford Tagger (RP) (Santorini 1990: 9-10).	Finite verbs	Le Foll
Verb features	*BE* -passives	**PASS**	*He must have been **burgled** . They need to be **informed** . He was **found** out. When were they **arrested** ?*	Assigned to past participles (here: VBN or VBD) preceded by the following patterns: 1) any form of the verb BE; 2) BE followed by one or two adverb(s) (RB) and/or a negation (XX0); 3) BE followed by a noun (NN.*) or personal pronoun (PRP); 4) BE followed by a noun (NN.*) or personal pronoun, and an adverb (RB) or negation (XX0). Unlike Biber (1988), no subdivision is made for by-passives and agentless passives. This choice is a) theoretically motivated because passives are too infrequent to be robustly measured at this level of granularity in most texts and b) for practical reasons because the algorithm proposed to identify by-passives resulted in too many false positives (e.g., looking for things that have been made by hand).	Finite verbs	Le Foll
Verb features	*GET*- passives	**PGET**	*He's gonna get **sacked.** She'll get me **executed** . It gets **done** all the time.*	Assigned to past participles (here: VBN or VBD) preceded by the following patterns: 1) any form of the verb GET; 2) GET followed by a noun (NN.*) or personal pronoun (PRP); 3) GET followed by a determiner (DT) or a noun (NN.*) plus a noun (NN.*).	Finite verbs	Le Foll

Verb features	*Going to* constructions	**GTO**	*I'm not **gonna** go. You're **going** to absolutely love it there! **Gonna** come along?*	Assigned to all occurrences of *going to* and *gonna* followed by a base form verb (VB), allowing for up to one intervening word between *going to* or *gonna* and the infinitive. GTO constructions are excluded from the progressive (PROG) count.	Finite verbs	Le Foll
Verb features	Past tense	**VBD**	*It **fell** and **broke**. I **implemented** it. If I **were** rich.*	As tagged by the Stanford Tagger, except where VBD tags are assumed to have been misassigned by the Stanford Tagger and are instead attributed to the perfect aspect (PEAS), passives (PASS, PGET) or USEDTO categories.	Finite verbs	Le Foll
Verb features	Non-finite verb *-ing* forms	**VBG**	*He texted me **saying** no. He just started **laughing**. I remember **thinking** about that.*	All verb forms ending in *-ing* as tagged by the Stanford Tagger, except those identified as progressives (PROG) or *going to* constructions (GTO). This category also includes "putative prepositions" ending in *-ing* such as *<u>according</u> to* and *<u>concerning</u> your request* (Santorini 1990: 11).	Finite verbs	Le Foll
Verb features	Non-finite *-ed* verb forms	**VBN**	*These include cancers caused by viruses. Our content is grouped into sections called topics. Have you read any of the books mentioned in the blog?*	As tagged by the Stanford Tagger except for the exclusion of tokens identified as instances of the perfect aspect (PEAS), passives (PASS, PGET) and *used to* constructions (USEDTO). Note that according to the Stanford Tagger rules, this category includes "putative prepositions" ending in *-ed* such as *<u>granted</u> that* and *<u>provided</u> that* (Santorini 1990: 11).	Finite verbs	Le Foll
Verb features	Imperatives	**VIMP**	***Let** me know! **Read** the website and **write** the names of the characters. In groups, **share** your opinion. Always **do** as you're told!*	This tag is first assigned to any verb in base form (VB) occurring 1) immediately after a punctuation mark except a comma (e.g., Okay: do it!), an emoji or emoticon (EMO), a symbol (SYM), hashtag (HST), foreign word (FW) or a list marker (LS), or 2) after a punctuation mark and an adverb (e.g., 1A. Then practice the dialogue), unless the VB token is please or thank or has previously been identified as a DO auxiliary (DOAUX). In a second loop, the VIMP tag is assigned to VB verb tokens (except thank or please) when preceded by an imperative as identified above, with up to two optional intervening tokens, and the tokens and or or (e.g., Describe or draw, Listen carefully and repeat, Read the text and answer the questions). In addition, a number of verbs frequently found in instructions are listed as exceptions (e.g., Complete, Choose, Check) and are always assigned to this category when they are found at the beginning of a sentence regardless of their tag because these were found to be frequently erronouesly identified by the Stanford Tagger as nouns (NN).	Finite verbs	Le Foll
Verb features	Present tense	**VPRT**	*It'**s** ours. Who **does** n't love it? I **know**.*	Subsumes the VBP (present tense other than third-person singular) and VBZ (third-person singular present tense) tags assigned by the Stanford Tagger. The MFTE also corrects systematic errors in the Stanford Tagger output by adding VPRT tags in strings such as *I dunno* and *there's.*	Finite verbs	Le Foll, adapted from Nini (2014)

Verb features	Perfect aspect	**PEAS**	*Have you **been** on a student exchange? She'd already **seen** it. He has been **told** before. Is this the last novel you've **read**?*	Assigned to past participles (VBN, VBD) preceded by the following patterns: 1) any form of the verb HAVE; 2) HAVE followed by one or two adverb(s) (RB) and/or a negation (XX0); 3) HAVE followed by a noun (NN.*) or personal pronoun (PRP); 4) HAVE followed by a noun (NN.*) or personal pronoun, and an adverb (RB) or negation (XX0); 5) HAVE followed by a participle tagged as a passive (PASS); 6) HAVE followed by one or two adverb(s) (RB) and/or a negation (XX0), and a passive participle (PASS); 7) HAVE followed by a noun (NN.*) or personal pronoun (PRP), and a passive participle (PASS); 8) *'s* as a verb (VBZ) followed by *been, had, done* or a stative verb; 9) 's as a verb (VBZ) followed by an adverb (RB) or negation (XX0), and *been, had, done* or a stative verb (as listed under JJPR).	Finite verbs	Le Foll
Verb features	Progressive aspect	**PROG**	*He wasn't **paying** attention. I'm **going** to the market. I'm **guessing** you're not going to be alone. I must be **getting** home.*	Assigned to any form of BE followed by an *-ing* form of any verb (VBG). The algorithm allows for an intervening adverb (RB), emphatic (EMPH) and/or negation (XX0). The interrogative form is captured as BE followed by a noun (N.*) or personal pronoun (PRP) followed by the VBG token. As for the affirmative version, the latter algorithm also accounts for an intervening adverb (RB) and/or negation (XX0). *Going to* constructions are excluded from this category and are tagged separately (GTO).	Finite verbs	Le Foll
Verb features	*HAVE got* constructions	**HGOT**	*He's **got** some. I haven't **got** any.*	Assigned to the word *got* preceded by the following patterns: 1) any form of the verb HAVE; 2) HAVE followed by one or two adverb(s) (RB) and/or a negation (XX0); 3) HAVE followed by a noun (NN, NNP) or personal pronoun (PRP); 4) HAVE followed by a noun (NNP, NNP) or personal pronoun, and an adverb (RB) or negation (XX0). Note that this algorithm overwrites the perfect aspect (PEAS) and passive (PASS) tag.	Finite verbs	Le Foll
Verb semantics	DO auxiliary	**DOAUX**	*Should take longer than it **does** . Ah you **did** . She needed that house, **did** n't she? You **don** 't really pay much attention, **do** you? Who **did** not already love him.*	Assigned to do, does and did as verbs in the following patterns: (a) when the next but one token is a base form verb (VB) (e.g., *did it work?, didn't hurt?*); (b) when the next but two token (+3) is a base form verb (VB) (e.g., *didn't it work)* ; (c) when it is immediately followed by an end-of-sentence punctuation mark (e.g., *you did?*); (d) when it is followed by a personal pronoun (PRP) or *not* or *n't* (XX0) and an end-of-sentence punctuation mark (e.g., *do you? He didn't!*); (e) when it is followed by *not* or *n't* (XX0) and a personal pronoun (PRP) (e.g., *didn't you?*); (f) when it is followed by a personal pronoun followed by any token and then a question mark (e.g., *did you really? did you not?*); (g) when it is preceded by a WH question word. Additionally, all instances of DO immediately preceded by *to* as an infinitive marker (TO) are excluded from this tag.	Finite verbs	Le Foll
Verb semantics	Activity verbs	**ACT**	*I got up and **ran** out. **Bring** your CV. Where have you **worked** before? I **go** to school.*	Assigned to all forms of the verbs: *buy, make, give, take, come, use, leave, show, try, work, move, follow, put, pay, bring, meet, play, run, hold, turn, send, sit, wait, walk, carry, lose, eat, watch, reach, add, produce, provide, pick, wear, open, win, catch, pass, shake, smile, stare, sell, spend, apply, form, obtain, arrange, beat, check, cover, divide, earn, extend, fix, hang, join, lie, obtain, pull, repeat, receive, save, share, smile, throw, visit, accompany, acquire, advance, behave, borrow, burn, clean, climb, combine, control, defend, deliver, dig, encounter, engage, exercise, expand, explore* and *reduce* (cf. Biber 2006: 246, based on the LGSWE, pp. 361–362, 367–368, 370). *Do* is only included when it has not previously been tagged as an auxiliary (DOAUX). *Get* and *go* were removed from Biber's (2006) list due to their high polysemy. Like Biber (2006), for practical reasons, no phrasal verbs were included in this variable.	Finite verbs	Le Foll, based on Biber (2006)

Verb semantics	Aspectual verbs	**ASPECT**	*You should just **keep** talking. I **started** early today.*	Following Biber (2006: 247, based on the LGSWE, pp. 364, 369, 371), assigned to all forms of the verbs: *start, keep, stop, begin, complete, end, finish, cease* and *continue*.	Finite verbs	Biber 2006
Verb semantics	Facilitation and causative verbs	**CAUSE**	*He **helped** her escape. I pleaded with her to **let** me go.*	Following Biber (2006: 247, based on the LGSWE, pp. 363, 369, 370), assigned to all forms of the verbs: *help, let, allow, affect, cause, enable, ensure, force, prevent, assist, guarantee, influence, permit* and *require*.	Finite verbs	Biber 2006
Verb semantics	Communication verbs	**COMM**	***Describe** it to your partner and **say** why. **Write** a list. **Say** what these words mean.*	Following Biber (2006: 247, based on the LGSWE, pp. 362, 368, 370), assigned to all forms of the verbs: *say, tell, call, ask, write, talk, speak, thank, describe, claim, offer, admit, announce, answer, argue, deny, discuss, encourage, explain, express, insist, mention, offer, propose, quote, reply, shout, sign, sing, state, teach, warn, accuse, acknowledge, address, advise, appeal, assure, challenge, complain, consult, convince, declare, demand, emphasize, excuse, inform, invite, persuade, phone, pray, promise, question, recommend, remark, respond, specify, swear, threaten, urge, welcome, whisper* and *suggest.* British spellings and the verbs *agree, assert, beg, confide, command, disagree, object, pledge, pronounce, plead, report, testify, vow* and *mean* were added. The latter was on Biber's (2006) list for mental verbs but, in most contexts encountered in the present study, it was found to be more likely to be a communication verb.	Finite verbs	Le Foll, based on Biber (2006)
Verb semantics	Existential or relationship verbs	**EXIST**	*Weren't they **representing** Jamaica? It encouraged young athletes to **stay** .*	Following Biber (2006: 247, based on the LGSWE, pp. 364, 369, 370–371), assigned to all forms of the verbs: *seem, stand, stay, live, appear, include, involve, contain, exist, indicate, concern, constitute, define, derive, illustrate, imply, lack, owe, own, possess, suit, vary, deserve, fit, matter, reflect, relate, remain, reveal, sound, tend* and *represent* . This variable does not include the copular *be* . *Look* was removed from Biber's original list because it frequently acts as an activity verb, too, e.g., *I was looking for my glasses.*	Finite verbs	Le Foll, based on Biber (2006)
Verb semantics	Mental verbs	**MENTAL**	*We **want** to **see** you tomorrow. Did you never **hear** back? I don't **recognize** any.*	Following Biber (2006: 246-247, based on the LGSWE, pp. 362–363, 368–369, 370), assigned to all forms of the verbs: *see, know, think, want, need* (unless identified as a necessity modal; cf. MDNE), *feel, like, hear, remember, believe, read, consider, suppose, listen, love, wonder, understand, expect, hope, assume, determine, agree, bear, care, choose, compare, decide, discover, doubt, enjoy, examine, face, forget, hate, identify, imagine, intend, learn, mind, miss, notice, plan, prefer, prove, realize, recall, recognize, regard, suffer, wish, worry, accept, appreciate, approve, assess, blame, bother, calculate, conclude, celebrate, confirm, count, dare, detect, dismiss, distinguish, experience, fear, forgive, guess, ignore, impress, interpret, judge, justify, observe, perceive, predict, pretend, reckon, remind, satisfy, solve, study, suspect* and *trust.* British spellings were added. *Afford* and *find* which can be found on Biber's original list, were removed for being too polysemous. Note that the phrase *dunno* , which is incorrectly parsed by the Stanford Tagger, was also retagged as *du_VPRT n_XX0 no_VB* and that *no_VB* tokens are also assigned to this category.	Finite verbs	Le Foll, based on Biber (2006)

Verb semantics	Occurrence verbs	**OCCUR**	*Couldn't have* ***happened*** *at a busier time! The cricket* ***lasts*** *all day.*	Following Biber (2006: 247, based on the LGSWE pp. 364, 369, 370), assigned to all forms of the verbs: *become, happen, change, die, grow, develop, arise, emerge, fall, increase, last, rise, disappear, flow, shine, sink, slip* and *occur* .	Finite verbs	Biber 2006
Verb semantics	Necessity modals	**MDNE**	*I really* ***must*** *go.* ***Should*** *n't you be going now? You* ***need*** *not have worried. Everybody* ***needed*** *to be needed.*	As in Biber (1988), all occurrences of *ought, should* and *must.* Contrary to Nini's operationalisation (2014: 27), only occurrences tagged as modals (MD) by the Stanford Tagger were included. In addition, *need* when tagged as a modal by the Stanford Tagger (mostly when followed by *not* or *n't*) or when immediately followed by *to* not tagged as a preposition (IN) was also added to this variable.	Finite verbs	Le Foll, adapted from Biber (1988)
Verb semantics	Modal *can*	**MDCA**	***Can*** *I give him a hint? You* ***can*** *not. I* ***ca*** *n't believe it!*	All occurrences of *can* and c*a* tagged as modals by the Stanford Tagger (MD). *Ca* was included because the Stanford Tagger parses *can't* as *ca + n't.*	Finite verbs	Le Foll
Verb semantics	Modal *could*	**MDCO**	*Do you think someone* ***could*** *have killed her? Well, that* ***could*** *be the problem.* ***Could*** *you do it by Friday?*	All occurrences of *could* tagged as a modal by the Stanford Tagger (MD).	Finite verbs	Le Foll
Verb semantics	Modals *may* and *might*	**MDMM**	***May*** *I have a word with you? But it* ***might*** *not be enough.*	All occurrences of *may* and *might* tagged as modals by the Stanford Tagger (MD).	Finite verbs	Le Foll
Verb semantics	*will* and *shall* modals	**MDWS**	*It* ***wo*** *n't do. Yes it* ***will*** *.* ***Shall*** *we see?*	The tokens *will* and *shall* and their contractions *'ll, wo* and *sha* when tagged as modals by the Stanford Tagger (MD).	Finite verbs	Le Foll
Verb semantics	modal *would*	**MDWO**	***Would*** *n't you like to know? If I could afford to buy it I* ***would*** *. I* ***'d*** *like to think it works.*	The tokens *will* and *shall* and their contractions *'ll, wo* and *sha* when tagged as modals by the Stanford Tagger (MD).	Finite verbs	Le Foll
Verb semantics	*be able to*	**ABLE**	*It should be* ***able*** *to speak back to you. Would you be* ***able*** *to?*	Assigned to occurrences of the bigram (un)able to, whenever (un)able has previously been identified as a predicative adjective (JJPR). These occurrences of (un)able are subsequently excluded from the JJPR count.	Finite verbs	Le Foll
				Tags not counted by MFTE but important to understand the operationalisation of other features		
Lexis	Foreign words	**FW**	*I chose turkish delight and* ***panna cotta . Merrry christmasss! Yo im*** *gonna love it!*	All remaining words tagged by the Stanford Tagger as foreign words and not identified as other variables by the MFTE. Frequently includes words spelt with non-standard spellings, missing apostrophes, and poorly OCR'ed due to unusual fonts. Note that this feature is not counted by the MFTE.	NA	Stanford Tagger
Lexis	Symbols	**SYM**	***â*** *2 € a go. I hope so* ***†*** *. That's* ******* *all* ******* *they said!*	All remaining non alphanumeric tokens tagged by the Stanford Tagger as symbols (SYM) or list markers (LS) and not identified as other variables by the MFTE. Also frequently includes words poorly OCR'ed due to unusual fonts or poorly encoded text. Note that this feature is not counted by the MFTE.	NA	Stanford Tagger

Verb features	*to* -infinitives	**TO**	*They were trying **to** find a solution. We like **to** think it's doable. I went in there **to** kinda like celebrate.*	Following Nini (2014: 21), all occurrences of *to* except when followed by another _IN token, a number (CD), determiner (DT), adjective (J.*), possessive pronoun (PRPS), WH-word (WPS, WDT, WP, WRB), pre-determiner (PDT), noun (N.*) or pronoun (PRP). Note that, unlike Nini (2014), this feature is only used to identify other linguistic features. All occurrences of *to* are counted as prepositions (IN) in the MFTE output tables.	NA	Nini (2014)
Verb features	Verb base form	**VB**	*She would **sit** and **read** most afternoons. What do you **use** it for? Ask your parents to **drive** you to your friend's house.*	As tagged by the Stanford Tagger, except those identified as imperatives (VIMP). This feature is not included in the tables of counts outputted by the MFTE because it overlaps with other features (e.g., all the modal verb features). However, it is used to identify many other linguistic features.	NA	Le Foll
Verb semantics	Private verbs	NA	*I don't **think** this should be **assumed**. I **suspect** he can't even **remember** it.*	As in Biber (1988, based on 1985: 1181), all forms of the verbs accept, *anticipate, ascertain, assume, believe, calculate, check, conclude, conjecture, consider, decide, deduce, deem, demonstrate, determine, discern, discover, doubt, dream, ensure, establish, estimate, expect, fancy, fear, feel, find, foresee, forget, gather, guess, hear, hold, hope, imagine, imply, indicate, infer, insure, judge, known, learn, mean, note, notice, observe, perceive, presume, presuppose, pretend, prove, realize, reason, recall, reckon, recognize, reflect, remember, reveal, see, sense, show, signify, suppose, suspect, think* and *understand.* Note that this category is only used to identify *that*- omissions (THATD).	NA	Biber 1988
Verb semantics	Public verbs	NA	*She **promised** she'd **write** back.*	As in Biber (1988, based on 1985: 1181), all forms of the verbs *acknowledge, add, admit, affirm, agree, allege, announce, argue, assert, bet, boast, certify, claim, comment, complain, concede, confess, confide, confirm, contend, convey, declare, deny, disclose, exclaim, explain, forecast, foretell, guarantee, hint, insist, maintain, mention, object, predict, proclaim, promise, pronounce, prophesy, protest, remark, repeat, reply, report, retort, say, state, submit, suggest, swear, testify, vow, warn* and *write.* Note that this category is only used to identify *that* -omissions (THATD).	NA	Le Foll, adapted from Biber (1988)
Verb semantics	Suasive verbs	NA	*They were **determined** to make this work. I'd **prefer** to do it that way.*	As in Biber (1988, based on 1985: 1182–3), all forms of the verbs *agree, allow, arrange, ask, beg, command, concede, decide, decree, demand, desire, determine, enjoin, ensure, entreat, grant, insist, instruct, intend, move, ordain, order, pledge, pray, prefer, pronounce, propose, recommend, request, require, resolve, rule, stipulate, suggest, urge* and *vote* . Note that this category is only used to identify *that* -omissions (THATD).	NA	Biber 1988
				Features removed from the MFTE feature portfolio post-evaluation of v.2.9 (Note that the corresponding lines are commented out in v.3.0+ and may still be run, if wished)		
Verb semantics	Quotative BE + *like*	**QLIKE**	*I was **like** oh this is really good. And everyone is **like** let's do this.*	Assigned to any form of *BE* followed by *like* tagged as a preposition (IN) by the Stanford Tagger and not followed by a noun (NN.*), adjective (J.*), determiner (DT), preposition (IN) or a full stop, comma, exclamation or question mark. This feature is deactivated by default but can be uncommented in the script.	Finite verbs	Le Foll

Verb features	*Used to* constructions	**USEDTO**	*You'll get **used** to it in time. It works but not like it **used** to .*	Assigned to all occurrences of the bigram *used to* . These occurrences of *used* are excluded from the VBN/VBD counts. This feature is deactivated by default but can be uncommented in the script.	Finite verbs	Le Foll
Discourse organisation	Phrasal coordination	**PHC**	*read **and** write, positive **or** negative, nouns **and** adjectives*	All occurrences of *and, &, or* and *nor* in the following patterns: adverb + *and/or/nor* + adverb, adjective + *and/or/nor* + adjective, verb + *and/or/nor* + verb, noun + *and/or/nor* + noun. This feature was removed post-evaluation and these occurrences of *and* and *or* are now all included in coordinating conjunctions (CC).	NA	Biber 1988

Versicherung an Eides statt über die Eigenständigkeit der erbrachten wissenschaftlichen Leistung[63]

Ich versichere hiermit an Eides statt, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Weitere Personen oder Organisationen waren an der inhaltlichen materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten, Promotionsberaterinnen oder Promotionsberatern oder anderen Personen in Anspruch genommen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Osnabrück, am 7. März 2022 ..... (Ort, Datum)

Elen Ge Foll

(Unterschrift)

---

[63] Nach § 9 Absatz 3 Satz 3, § 7 Absatz 4 Satz 2 NHG darf die Universität von den Doktorandinnen und Doktoranden eine Versicherung an Eides statt verlangen und abnehmen, wonach die Promotionsleistung von ihnen selbständig und ohne unzulässige fremde Hilfe erbracht worden ist.
Die Abgabe einer falschen eidesstattlichen Versicherung ist strafbar. Bei vorsätzlicher, also wissentlicher, Abgabe einer falschen Erklärung droht eine Freiheitsstrafe bis zu 3 Jahren oder eine Geldstrafe. Eine fahrlässige Abgabe (obwohl hätte erkannt werden müssen, dass die Erklärung nicht den Tatsachen entspricht) kann eine Freiheitsstrafe bis zu einem Jahr oder eine Geldstrafe nach sich ziehen.
§ 156 StGB: Falsche Versicherung an Eides Statt
Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.
§ 161 StGB: Fahrlässiger Falscheid, fahrlässige falsche Versicherung an Eides Statt:
(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.
(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.