Vers une analyse des différences interlinguistiques entre les genres textuels : étude de cas basée sur les n-grammes et l'analyse factorielle des correspondances

Marie-Aude Lefer, Yves Bestgen, Natalia Grabar

Vers une analyse des différences interlinguistiques entre les genres textuels : étude de cas basée sur les n-grammes et l'analyse factorielle des correspondances

Marie-Aude Lefer[1, 2] Yves Bestgen[2] Natalia Grabar[3]
(1) Université Saint-Louis, Bruxelles, Belgique
(2) Université catholique de Louvain, Louvain-la-Neuve, Belgique
(3) CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
marie-aude.lefer@uclouvain.be, yves.bestgen@uclouvain.be,
natalia.grabar@univ-lille3.fr

RÉSUMÉ

L'objectif de notre travail est d'évaluer l'intérêt d'employer les n-grammes et l'analyse factorielle des correspondances (AFC) pour comparer les genres textuels dans les études contrastives interlinguistiques. Nous exploitons un corpus bilingue anglais-français constitué de textes originaux comparables. Le corpus réunit trois genres : les débats parlementaires européens, les éditoriaux de presse et les articles scientifiques. Dans un premier temps, les n-grammes d'une longueur de 2 à 4 mots sont extraits dans chaque langue. Ensuite, pour chaque longueur, les 1 000 n-grammes les plus fréquents dans chaque langue sont traités par l'AFC pour déterminer quels n-grammes sont particulièrement saillants dans les genres étudiés. Enfin, les n-grammes sont catégorisés manuellement en distinguant les expressions d'opinion et de certitude, les marqueurs discursifs et les expressions référentielles. Les résultats montrent que les n-grammes permettent de mettre au jour des caractéristiques typiques des genres étudiés, de même que des contrastes interlangues intéressants.

ABSTRACT

Towards a cross-linguistic analysis of genres: A case study based on n-grams and Correspondence Analysis

The aim of the present study is to assess the use of n-grams and Correspondence Analysis (CA) to compare genres in cross-linguistic studies. The study is based on an English-French bilingual corpus made up of original (*i.e.* non-translated) texts, representing three genres: European parliamentary debates, newspaper editorials and academic articles. First, 2- to 4-grams are extracted in each language. Second, the most frequent 1000 n-grams for each n-gram length and in each language are analyzed by means of CA with a view to determining which n-grams are particularly salient in the genres examined. Finally, n-grams are manually classified into a range of categories, such as stance expressions, discourse markers and referential expressions. The results show that the n-gram approach makes it possible to uncover typical features of the three genres investigated, as well as interesting contrasts between English and French.

MOTS-CLÉS : corpus comparables, analyse factorielle des correspondances, n-grammes, genres textuels.

KEYWORDS: Comparable Corpora, Correspondence Analysis, N-grams, Genres.

1 Introduction et Objectifs

Parmi les études comparatives, on distingue typiquement les études contrastives (c'est-à-dire inter-linguistiques), dans lesquelles des travaux sur au moins deux langues sont effectués, et les études monolingues, dans lesquelles d'autres types de contrastes sont recherchés, comme par exemple l'étude de registres ou de genres textuels au sein d'une même langue (Biber, 1988). Il est également possible de combiner ces deux approches, à savoir comparer simultanément les langues et les genres textuels (Neumann, 2013, 2014; Biber, 2014). C'est ce que nous proposons de faire dans cette étude.

La majorité des travaux interlinguistiques basés sur corpus se limitent à l'étude d'un seul genre, essentiellement les romans ou la presse journalistique, et ont tendance à généraliser les résultats obtenus aux systèmes langagiers comparés, comme s'il s'agissait d'entités monolithiques (Lefer & Vogeleer, 2014). Au vu des limites de ce type d'approche, des travaux contrastifs plus récents, dédiés à un phénomène linguistique particulier (qu'il soit lexical, syntaxique, discursif, etc.), prennent en compte les variations d'un genre à l'autre (Granger, 2014; Hasselgård, 2014; Gómez González, 2014). Ces études montrent clairement que certains contrastes interlangues ne sont en fait avérés que dans certains genres textuels, et ne sont donc pas généralisables à l'ensemble de la langue. Á cet égard, une approche novatrice en linguistique contrastive, proposée par Neumann (2013, 2014), consiste à prendre les genres textuels comme objet d'étude à proprement parler, en les comparant dans plusieurs langues à la fois. Neumann étudie différents genres en anglais et en allemand au travers d'un ensemble d'indicateurs linguistiques tirés de la grammaire fonctionnelle systémique, comme les pronoms personnels, les verbes modaux, les nominalisations, etc. On peut également citer l'analyse multi-dimensionnelle proposée par Biber, qui a été employée pour étudier la variation entre les genres dans différentes langues (prises séparément), comme l'espagnol, le portugais, le tchèque, le coréen, etc. (voir (Biber, 2014) pour un aperçu général de ces études).

Dans ce contexte, nous proposons d'évaluer une autre méthode d'analyse interlinguistique des genres textuels, en ayant recours aux n-grammes de mots comme unique indicateur linguistique. L'intérêt d'étudier les n-grammes est double. D'une part, il s'agit d'une démarche «corpus-driven» ou guidée par le corpus (Biber, 2009; Cortes, 2015), qui ne se base pas a priori sur un inventaire d'indicateurs jugés pertinents par l'analyste, une démarche nécessairement subjective, mais découvre ces indicateurs par une analyse automatique du corpus. D'autre part, les n-grammes permettent de mettre au jour une large gamme d'objets linguistiques de différents types, comme par exemple les expressions référentielles, les marqueurs discursifs et les expressions d'opinion et de certitude (« stance » en anglais) (Biber *et al.*, 2004), de même que d'autres types de séquences récurrentes qui jouent des rôles importants dans la langue (*e.g.* structures passives, certains types de syntagmes nominaux). En d'autres mots, selon notre hypothèse, les n-grammes peuvent aider à mettre au jour des contrastes discursifs et rhétoriques importants entre les langues (Ebeling & Oksefjell Ebeling, 2013; Granger, 2014) et entre les genres (Biber, 1988).

Dès lors, l'objectif de notre étude est d'évaluer la possibilité et l'intérêt d'employer les n-grammes pour effectuer des études contrastives des genres textuels en utilisant l'analyse factorielle des correspondances ou AFC (Lebart *et al.*, 2000) pour faciliter une analyse linguistique détaillée des différences entre les genres et les langues. Cette technique, fréquemment employée en statistique textuelle, repose, comme nombre d'autres techniques de réduction de données employées en TAL, sur la décomposition en valeurs singulières. Nous l'avons choisie en raison (1) de son orientation vers la visualisation des données et (2) des différents indices numériques d'aide à l'interprétation des résultats qu'elle fournit, qui se révèlent fort utiles lors de l'analyse linguistique fine. Cette technique

est statistiquement beaucoup plus adéquate pour traiter des tables de contingence que, par exemple, l'analyse en composante principale fréquemment employée en linguistique appliquée (p.e., (Biber, 1988)). Différentes études consacrées aux genres textuels ont eu recours à l'AFC. On peut par exemple citer les études de textes littéraires (Brunet, 2004, 2003), tirés du web (Habert *et al.*, 2000), politiques (Habert, 1983) ou encore idéologiques (Valette & Grabar, 2004). Allant vers un niveau plus fin d'analyse, l'AFC a également été utilisée pour étudier la morphologie (Brunet, 1981), le lexique (Habert, 1983; Brunet, 1999), les segments répétés, qui est une autre appellation pour les n-grammes de mots (Salem, 1984) ou encore la similarité sémantique entre des phrases (Bestgen, 2014). Dans la suite de cet article, nous décrivons d'abord les données analysées et la méthode proposée (section 2). Ensuite, nous présentons et discutons les résultats (section 3) et concluons avec les perspectives de ce travail (section 4).

2 Données et Méthode

Nous exploitons trois corpus bilingues comparables, c'est-à-dire des corpus comprenant uniquement des textes originaux en anglais et en français (et non des textes traduits) :

– le corpus Europarl, qui contient les transcriptions des débats du Parlement européen (Koehn, 2005; Cartoni & Meyer, 2012) ;
– le corpus KIAP, qui contient des articles scientifiques issus de trois domaines (médecine, économie et linguistique) (Fløttum *et al.*, 2006) ;
– le corpus Multed, qui contient des éditoriaux de la presse journalistique [1].

Chaque corpus comprend 600 000 mots par langue.

Notre méthode comporte plusieurs étapes :

1. *Préparation des corpus*. Les corpus ont été divisés dans chaque langue en quatre tranches homogènes de 150 000 mots chacune. Cela correspond à :
   (a) quatre ans de débats parlementaires (1996 à 1999) ;
   (b) quatre sous-corpus disciplinaires issus de KIAP: un sous-corpus médical, deux sous-corpus économiques et un sous-corpus linguistique dans chaque langue [2] ;
   (c) quatre sous-corpus journalistiques représentant chacun un quotidien différent (par exemple, pour le français, Le Monde, Libération, Le Figaro et Le Nouvel Observateur).
2. *Extraction des n-grammes*. Les n-grammes d'une longueur de 2 à 4 mots sont extraits de l'ensemble des corpus étudiés. Les 1 000 n-grammes les plus fréquents pour chaque longueur (2 à 4 mots) et dans chaque langue sont retenus. Il est à noter que les n-grammes les plus longs sont fréquemment composés de n-grammes plus courts eux-mêmes sélectionnés pour être analysés (par exemple, *ce qui concerne*$_{3-gramme}$, *en ce qui concerne*$_{4-gramme}$, *pour ce qui concerne*$_{4-gramme}$). Ceci ne pose toutefois pas de problèmes puisque les AFC sont effectuées séparément pour chaque longueur de n-grammes.
3. *Analyse factorielle des correspondances*. Pour chaque longueur de n-grammes et chaque langue, le tableau de contingence qui est formé en croisant les fréquences des n-grammes dans les sous-corpus est soumis à une analyse factorielle des correspondances. Cette procédure statistique,

1. https://www.uclouvain.be/en-cecl-multed.html
2. Le corpus d'articles scientifiques utilisé ici comporte deux sous-corpus économiques car il s'agit du seul sous-corpus qui atteint au moins 300 000 mots dans KIAP.

disponible dans de nombreux logiciels comme R (Nenadic & Greenacre, 2007), décompose le tableau en une série de dimensions orthogonales ordonnées selon la part des associations entre les lignes (les n-grammes) et les colonnes (les sous-corpus) qu'elles expliquent. Chaque n-gramme et chaque sous-corpus se voit affecter des coordonnées sur ces dimensions et ces coordonnées sont utilisées pour représenter graphiquement les distances entre les sous-corpus et entre les n-grammes. Cette représentation graphique simultanée [3] des lignes et des colonnes du tableau lexical facilite grandement l'interprétation des dimensions.

3 Résultats

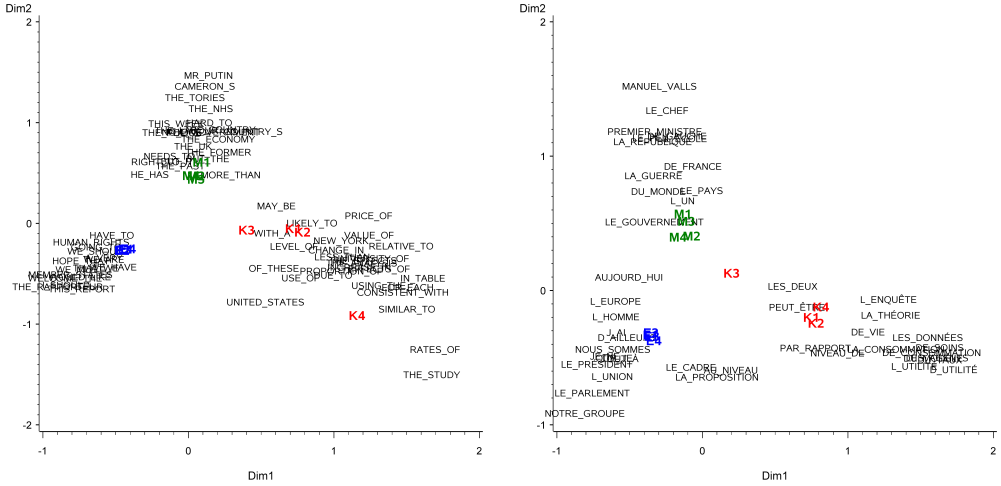

Fig. 1 – Nuages de 2-grammes en fonction des genres (anglais à gauche et français à droite) pour les deux premières dimensions. Note: E est employé pour désigner le corpus *Europarl*, M pour *Multed* et K pour *KIAP*

Comme les corpus ont été sélectionnés dans chaque langue de manière à être aussi comparables que possible, on ne s'étonnera pas de la grande ressemblance entre les résultats pour les deux langues tant dans les parts d'inertie [4] expliquées par les premières dimensions que dans les sous-corpus que celles-ci opposent. Par exemple, pour les 2-grammes, deux dimensions sur les 11 possibles expliquent plus de la moitié de l'inertie, soit presque trois fois plus que ce à quoi l'on s'attendrait s'il y avait une totale indépendance entre les sous-corpus et les n-grammes, et leur interprétation est la même dans les deux langues. Dans la figure 1, où seule une partie des 2-grammes est indiquée afin d'en améliorer la lisibilité, on remarque par exemple que la première dimension trouvée par l'AFC oppose les articles scientifiques aux débats parlementaires alors que la deuxième dimension oppose les éditoriaux de presse aux autres genres, et ce dans les deux langues. Dans les dimensions suivantes, les domaines

3. Pour rappel, il n'est pas légitime d'interpréter directement la proximité entre un point-ligne et un point-colonne. Il s'agit plutôt d'employer ces deux sources d'information pour interpréter les dimensions.
4. Notion similaire à celle de variance dans une analyse factorielle.

représentés dans le corpus KIAP sont bien différenciés. Des résultats semblables sont obtenus pour les 3- et 4-grammes.

Par ailleurs, lorsque tous les n-grammes distinctifs sont catégorisés manuellement, en suivant une typologie inspirée en partie des travaux de Biber (Biber *et al.*, 2004), on remarque des similarités fortes dans les deux langues. L'analyse détaillée des résultats obtenus grâce à l'analyse factorielle des correspondances indique que :

– les articles scientifiques sont bien différenciés des débats parlementaires. En effet, il ressort de l'analyse qualitative des données que les articles scientifiques privilégient l'utilisation de substantifs post-modifiés par un syntagme prépositionnel (*e.g.*, {*the difference between*; *les résultats de*}) et de structures existentielles introduites par {*there*; *il*} (*e.g.*, {*there were no*; *il existe un*}), tandis que les débats parlementaires contiennent beaucoup d'expressions d'opinion ou de certitude avec un pronom personnel à la première personne (*e.g.*, {*I hope that*; *je crois que*}) ;
– les éditoriaux de presse s'opposent aux autres corpus, de par la présence de nombreux noms propres (*e.g.*, *the Liberal Democrats/François Hollande*) et d'expressions de temps et de lieu (*e.g.*, *in recent years/sur la scène internationale*) ;
– au sein du corpus KIAP, il existe une différence nette entre les articles en médecine et en économie, ce qui est dû principalement aux expressions référentielles (termes complexes) (*e.g.*, *traitement de substitution/taux de croissance*).

Même si les dimensions de la variation entre les genres sont très proches dans les deux langues, il apparaît que deux types d'unités phraséologiques, les marqueurs discursifs et les expressions d'opinion et de certitude, jouent un rôle plus central en anglais pour la distinction entre les genres. En revanche, en français, les résultats indiquent que ces unités contribuent moins à la construction des dimensions. La figure 2, qui ne reprend que les marqueurs discursifs et les expressions d'opinion et de certitude, illustre cette tendance dans les éditoriaux de presse pour les 4-grammes.

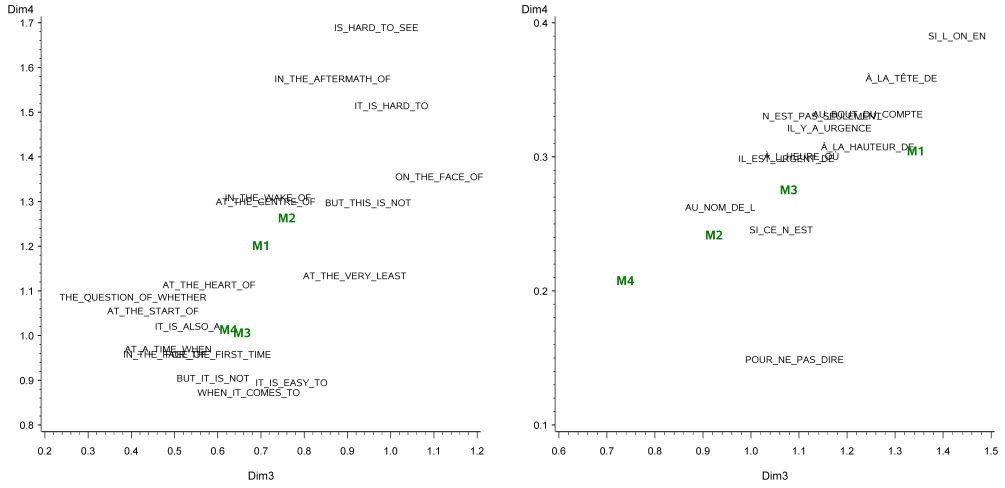

Fig. 2 – Nuages de 4-grammes en fonction des genres (anglais à gauche et français à droite) pour les dimensions 3 et 4. Note: seuls les corpus *Multed* et deux types d'unités phraséologiques sont représentés

Ce contraste entre l'anglais et le français peut être relié à des différences systémiques entre les deux

langues, le français étant réputé comme ayant plus fréquemment recours aux marqueurs discursifs et aux expressions d'opinion et de certitude (e.g. (Vinay & Darbelnet, 1995)), et ce, peu importe le genre textuel. En d'autres termes, en français, ces n-grammes semblent être partagés par les genres, plutôt que typiques de tel ou tel genre, comme c'est le cas en anglais (Granger, 2014).

4 Conclusion et Perspectives

Cette recherche visait à évaluer la possibilité d'effectuer des études contrastives des genres textuels par une approche TAL de type «corpus-driven». Pour ce faire, nous avons comparé trois genres différents dans deux langues sur la base des n-grammes les plus fréquents en utilisant l'analyse factorielle des correspondances pour faciliter les analyses linguistiques détaillées. Les résultats nous semblent encourageants. Non seulement la méthode a corroboré nos hypothèses et mis en évidence, comme attendu, de grandes ressemblances entre les deux langues, mais elle a aussi permis de mettre au jour des différences interlinguistiques dans l'emploi d'unités phraséologiques comme les expressions d'opinion et de certitude.

Habituellement, ce genre d'études est mené en dressant dans chaque genre la liste des n-grammes les plus fréquents et en comparant manuellement ces listes. Afin que cette comparaison reste réalisable, les chercheurs limitent fréquemment leurs analyses aux 100 ou 200 expressions les plus fréquentes et surtout n'extraient des données qu'une information très rudimentaire : à combien de genres un n-gramme donné est-il commun ? Dans cette recherche, nous proposons d'extraire des informations nettement plus riches d'un nombre beaucoup plus grand de n-grammes, tout en conservant une approche «corpus-driven». Cette procédure non supervisée, qui permet d'acquérir les n-grammes, ou les expressions, les plus typiques de différents genres textuels dans deux langues, ouvre la voie à la création de lexiques bilingues contextualisés. De telles ressources sont extrêmement utiles en TAL pour différents contextes et applications, comme par exemple la traduction automatique et assistée par ordinateur, la formation de traducteurs, la préparation de descripteurs pour la recherche d'information et l'apprentissage supervisé.

Un problème majeur de ce type d'analyses comparatives (genres/langues) qu'il nous reste à aborder est que les n-grammes discriminatifs peuvent être dus à des différences thématiques entre les genres comparés (des mots lexicaux qui sont nécessairement différents). Il n'est cependant pas aisé de les supprimer en amont des analyses. En fait, c'est même discutable puisqu'ils participent à la distinction entre les genres et qu'il est fort probable que la manière dont certains d'entre eux sont employés (déterminant, construction syntaxique, etc.) est différente selon la langue ou le genre. L'idée est donc de les traiter en aval par une analyse linguistique fine, mais cela nécessite d'employer une technique d'analyse qui facilite cette démarche. Les observations rapportées ci-dessus suggèrent que l'AFC pourrait être cette technique. Analyser d'une manière beaucoup plus approfondie ces expressions est la principale piste que nous aimerions explorer à l'avenir. À terme, ces analyses pourraient permettre d'automatiser au moins partiellement l'identification de certaines catégories d'expressions. Dans nos travaux futurs, nous comptons également comparer les observations et les conclusions selon la longueur des n-grammes. En effet, plus ils sont longs, plus ils sont précis, mais d'un autre côté, les n-grammes longs sont plus rares. On peut dès lors se demander si les analyses linguistiques seront plus concluantes si on privilégie la fréquence ou la précision. Une réflexion similaire sera menée en ce qui concerne le degré idéal de lemmatisation des n-grammes (p.e., contrairement à *de, le* ou *que*, il n'est pas pertinent de fusionner les formes fléchies de *falloir*, comme *faut* et *faudrait*, qui

caractérisent des expressions d'opinion bien distinctes). En outre, il faudra étudier dans le corpus comment sont employés les n-grammes (non-thématiques) qui distinguent le plus les genres et dont le fonctionnement est différent selon la langue. Il serait également pertinent d'analyser d'autres langues afin de mieux comprendre la différence entre l'anglais et le français en ce qui concerne les marqueurs discursifs et les expressions d'opinion et de certitude. En d'autres termes, est-ce qu'une des deux langues analysées est différente de nombreuses autres (et si oui, laquelle) ou s'agit-il plutôt d'un continuum ?

Plus généralement, nous aimerions également comparer les résultats obtenus ici à l'aide des n-grammes et de l'AFC à d'autres types d'indicateurs linguistiques (comme les POS-grammes ou les motifs (Quiniou *et al.*, 2012; Longrée & Mellet, 2013)) et comparer l'AFC aux métriques de la recherche d'information (*tf-idf*) ou à d'autres méthodes statistiques comme l'analyse sémantique latente ou l'allocation latente de Dirichlet (Latent Dirichlet Allocation). Enfin, nous envisageons d'étendre nos travaux à l'analyse comparative de la langue originale et de la langue traduite (appelée «translationese», voir par exemple (Volansky *et al.*, 2014; Avner *et al.*, 2016; Bestgen & Granger, 2016)).

Remerciements

Yves Bestgen est chercheur qualifié du F.R.S.-FNRS.

Références

AVNER E., ORDAN N. & WINTNER S. (2016). Identifying translationese at the word and sub-word level. *Digital Scholarship in the Humanities*, **31**(1), 30–54.

BESTGEN Y. (2014). CECL: a new baseline and a non-compositional approach for the sick benchmark. In *8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 160–165, Dublin, Ireland.

BESTGEN Y. & GRANGER S. (2016). Collocation et traduction. analyse automatique au moyen d'indices d'association. In M. KAUFFER & Y. KEROMNES, Eds., *Theorie und Empirie in der Phraseologie / Approches théoriques et empiriques en phraséologie*. Tübingen: Stauffenburg Verlag.

BIBER D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

BIBER D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, **14**(3), 275–311.

BIBER D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. In M.-A. LEFER & S. VOGELEER, Eds., *Languages in Contrast. Special issue Genre- and Register-related Discourse Features in Contrast*, volume 14, p. 7–34. Amsterdam/Philadelphia: John Benjamins.

BIBER D., CONRAD S. & CORTES V. (2004). If you look at ... lexical bundles in university lectures and textbooks. *Applied Linguistics*, **25**, 371–405.

BRUNET E. (1981). *Les suffixes*, In *Le vocabulaire français de 1789 à nos jours. D'après les données du Trésor de la langue française*, p. 415–493. Librairie Slatkine.

BRUNET E. (1999). *Aperçu statistique sur l'évolution du vocabulaire français*, In *Nouvelle histoire de la langue française*, p. 675–627. Éditions du Seuil.

BRUNET E. (2003). *Nouvelles méthodes statistiques. L'exemple de Rabelais*, In *Ancien et moyen français sur le Web*, p. 33–54. Les éditions DAVID.

BRUNET E. (2004). Statistiques rimbaldiennes. In *Les littératures de l'Europe unie*, Cesenatico, Université de Bologne.

CARTONI B. & MEYER T. (2012). Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *8th International Conference on Language Resources and Evaluation (LREC)*.

CORTES V. (2015). Situating lexical bundles in the formulaic language spectrum: Origins and functional analysis developments. In V. CORTES & E. CSOMAY, Eds., *Corpus-based Research in Applied Linguistics: Studies in Honor of Doug Biber*, p. 197–216. Amsterdam: John Benjamins.

EBELING J. & OKSEFJELL EBELING S. (2013). *Patterns in Contrast*. Amsterdam & Philadelphia: Benjamins.

FLØTTUM K., DAHL T. & KINN T. (2006). *Academic Voices – across languages and disciplines*. Amsterdam & Philadelphia: Benjamins.

GRANGER S. (2014). A lexical bundle approach to comparing languages: Stems in English and French. In M.-A. LEFER & S. VOGELEER, Eds., *Languages in Contrast. Special issue Genre- and Register-related Discourse Features in Contrast*, volume 14, p. 58–72. Amsterdam/Philadelphia: John Benjamins.

GÓMEZ GONZÁLEZ M. (2014). Canonical tag questions in English, Spanish and Portuguese. a discourse-functional study. In M.-A. LEFER & S. VOGELEER, Eds., *Languages in Contrast. Special issue Genre- and Register-related Discourse Features in Contrast*, volume 14, p. 93–126. Amsterdam/Philadelphia: John Benjamins.

HABERT B. (1983). Études des formes spécifiques et typologie des énoncés (les résolutions générales des congrès de la CFTC-CFDT de 1945 à 1979). *MOTS, Presses de la Fondation Nationale des Sciences Politiques*, (7), 97–124.

HABERT B., ILLOUZ G., LAFON P., FLEURY S., FOLCH H., HEIDEN S. & PRÉVOST S. (2000). Profilage de textes : cadre de travail et expérience. In M. RAJMAN, Ed., *5èmes Journées d'Analyse des Données Textuelles (JADT)*, Lausanne.

HASSELGÅRD H. (2014). Discourse-structuring functions of initial adverbials in English and Norwegian news and fiction. In M.-A. LEFER & S. VOGELEER, Eds., *Languages in Contrast. Special issue Genre- and Register-related Discourse Features in Contrast*, volume 14, p. 73–92. Amsterdam/Philadelphia: John Benjamins.

KOEHN P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*, p. 79–86.

LEBART L., MORINEAU A. & PIRON M. (2000). *Statistique exploratoire multidimensionnelle (3ième édition)*. Paris: Dunod.

M.-A. LEFER & S. VOGELEER, Eds. (2014). *Genre- and Register-related Discourse Features in Contrast. Special issue of Languages in Contrast*. John Benjamins.

LONGRÉE D. & MELLET S. (2013). Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours. *Langages*, **189**, 68–80.

NENADIC O. & GREENACRE M. (2007). Correspondence analysis in r, with two- and three-dimensional graphics: the ca package. *Journal of Statistical Software*, **20**(3), 1–13.

NEUMANN S. (2013). *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin: de Gruyter Mouton.

NEUMANN S. (2014). Cross-linguistic register studies: Theoretical and methodological considerations. In M.-A. LEFER & S. VOGELEER, Eds., *Languages in Contrast. Special issue Genre- and Register-related Discourse Features in Contrast*, volume 14, p. 35–57. Amsterdam/Philadelphia: John Benjamins.

QUINIOU S., CELLIER P., CHARNOIS T. & LEGALLOIS D. (2012). Fouille de données pour la stylistique: l'exemple des motifs émergents. In *11es Journées Internationales d'analyse statistique des données textuelles*, p. 821–833, Liège, Belgium.

SALEM A. (1984). La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes. *Les Cahiers d'Analyse des Données*, **IX**(4), 489–500.

VALETTE M. & GRABAR N. (2004). Caractérisation de textes à contenus idéologiques : statistique textuelle ou extraction de syntagme ? L'exemple du projet PRINCIP. In *Journées de traitement automatique des données textuelles (JADT)*, Liège, Belgique.

VINAY J.-P. & DARBELNET J. (1995). *Comparative Stylistics of French and English. A Methodology for Translation*. Amsterdam & Philadelphia: Benjamins.

VOLANSKY V., ORDAN N. & WINTNER S. (2014). On the features of translationese. *Literary and Linguistic Computing*, p. 98–118.