Traitement des productions dans le corpus multimodal d'apprenants : AntConc au service de la recherche en linguistique de corpus

Evgenia Nicol-Bakaldina

DEMC Journal

N° 2 | 2024
2024

Traitement des productions dans le corpus multimodal d’apprenants : AntConc au service de la recherche en linguistique de corpus

Evgenia Nicol-Bakaldina
Docteur
Département des Etudes Anglophones, Faculté des Lettres et Sciences Humaines
CRILLASH
Université des Antilles , Fort-de-France

Édition électronique :
URL :
https://demc-journal.org/articles/revue-2/3742-traitement-des-productions-dans-le-corpus-multimodal-d-apprenants-antconc-au-service-de-la-recherche-en-linguistique-de-corpus
ISSN : 3036-5295
Date de publication : 15/02/2024

Pour **citer cette publication** : Nicol-Bakaldina, E. (2024). Traitement des productions dans le corpus multimodal d’apprenants : AntConc au service de la recherche en linguistique de corpus. *DEMC Journal*, (2). https://demc-journal.org/articles/revue-2/3742-traitement-des-productions-dans-le-corpus-multimodal-d-apprenants-antconc-au-service-de-la-recherche-en-linguistique-de-corpus

RÉSUMÉ. Cet article présente une méthodologie d'analyse de corpus multimodal des apprenants en secondaire dans le cadre d’une recherche doctorale sur l’enseignement bilingue EMILE, en utilisant des outils de traitement automatique de la langue (TAL), notamment AntConc. À travers une étude de cas, l'article met en lumière les fonctionnalités d’AntConc, telles que les requêtes simples, l’analyse de la fréquence et du type de lexique, la distribution des mots dans les productions orales et écrites, ainsi que la génération de lignes de concordance et la recherche de collocations. Ces fonctionnalités permettent d'explorer comment les apprenants utilisent le modèle linguistique proposé par leurs enseignants dans un contexte bilingue. L’article conclut en évaluant les avantages méthodologiques et les limites de l’utilisation d'AntConc dans l’analyse de corpus multimodaux, soulignant son importance pour le développement de nouvelles approches en linguistique appliquée.

ABSTRACT. This article presents a methodological approach for analyzing multimodal corpora within the framework of doctoral research on bilingual education (EMILE) in a secondary school, using language processing tools (TAL), in particular AntConc. Basing on a case study, the article highlights the key functionalities of AntConc, such as simple queries, the analysis of lexical frequency and type, word distribution in oral and written productions, as well as the generation of concordance lines and the search for collocations. These features help explore how learners utilize the linguistic models provided by their teachers in a bilingual context. The article concludes with an evaluation of the methodological advantages and limitations of using AntConc for multimodal corpus analysis, emphasizing its importance for the development of new approaches in applied linguistics.

Mots-clefs :
Corpus, EMILE, CLIL, TAL, École sécondaire, Vocabulaire, NLP tools, Vocabulary, Secondary school, Corpus de linguistique

---

Introduction et problématique

Le domaine de la linguistique computationnelle a connu un essor considérable depuis les années 2000, grâce à l'évolution des logiciels de traitement automatique de la langue (TAL). La liste non- exhaustive d’applications des outils TAL inclut la transcription, l’étiquetage, les recherches ciblées des collocations et d’usage des mots en contexte, les analyses morpho-syntaxiques, lexiques ou sémantiques, reconnaissance vocale, parmi plusieurs autres (voir Tutin et coll., s.d; Rohlfing et coll., 2006). En fonction de l’aspect envisagé du travail avec une langue, il est possible de parler de « ‘l’ingénierie linguistique’, quand l’accent est mis sur les aspects pratiques et

opérationnels, ou de 'linguistique informatique', quand c'est la linguistique qui tient un rôle important dans les recherches »[1]. Parmi ces outils, AntConc joue un rôle clé dans le développement de méthodologies appliquées à l'analyse des corpus multimodaux.

Notre contribution s'inscrit dans un cadre de développement méthodologique en linguistique appliquée. Il vise à démontrer comment l'utilisation d'outils TAL, notamment AntConc, permet d'analyser des corpus oraux et écrits issus de productions d'apprenants dans un contexte d'enseignement bilingue (EMILE)[2]. L'accent est mis sur la manière dont ces logiciels permettent d'explorer des aspects lexicaux et syntaxiques à partir de données multimodales tirées du corpus CORINE-SEHEA, contribuant ainsi à enrichir la recherche en didactique des langues. En effet, l'utilisation de logiciels tels qu'AntConc enrichit la recherche en didactique des langues en offrant des moyens efficaces aux professeurs pour analyser les productions des élèves sous un angle lexical et morpho-syntaxique. Ces analyses permettent notamment d'identifier les caractéristiques de l'interlangue, de suivre l'acquisition du vocabulaire et d'observer la réutilisation des structures enseignées, en particulier en contexte EMILE. Par ailleurs, en comparant l'input (ce qui est enseigné) des enseignants avec l'output (les productions) des élèves, AntConc met en lumière les écarts et les points d'ancrage pédagogiques, facilitant l'adaptation des supports et des méthodes d'enseignement. Le logiciel permet également d'évaluer l'impact de l'approche interdisciplinaire EMILE sur l'apprentissage linguistique, tout en ouvrant des perspectives méthodologiques innovantes grâce à l'intégration de données multimodales. Ainsi, AntConc contribue à une compréhension dynamique des processus d'acquisition et à l'amélioration des pratiques pédagogiques.

Le logiciel AntConc renseigne notamment sur les dimensions linguistiques de notre objet d'étude, notamment concernant les caractéristiques et les traits spécifiques de l'interlangue des élèves (vocabulaire, spécificités grammatico-lexicales, etc.). Nous nous interrogeons sur la place et la nature du vocabulaire dans l'apprentissage immersif EMILE : appartient-il davantage au registre académique, spécialisé, ou se situe-t-il à l'intersection des deux ? Quels sont les mots et les expressions les plus fréquents utilisés par les apprenants ? Le mot fréquent est-il employé plus souvent en cours de langue ou en cours de la DNL ? Le contexte d'emploi des mots par les élèves est-il le même que celui utilisé par les professeurs ? Enfin, que-est- ce que ces analyses révèlent sur le processus de l'apprentissage et sur l'interlangue des élèves en cours immersif EMILE ?

Dans la première partie, l'article esquisse le dispositif EMILE en Europe et en France, tout en expliquant que le lien entre les langues étrangères et le français repose sur une complémentarité éducative : les programmes français intègrent des cours pluridisciplinaires où le français sert souvent de support à l'apprentissage de langues étrangères et vice-versa, favorisant ainsi le développement d'une compétence plurilingue et une ouverture interculturelle conformément aux initiatives européennes. Le contexte de l'étude est présenté par la suite, notamment le profil des participants et le développement du projet. Le corpus d'étude CORINE est décrite ainsi que les éléments liés au recueil de donnés. Dans la partie méthodologique, nous décrivons le corpus des natifs anglophones LOCNESS (Louvain Corpus of Native English Essays), qui

a servi de corpus de référence pour établir une comparaison pertinente entre les productions des élèves français et anglophones, notamment dans le but d'évaluer l'utilisation du lexique spécialisé et de déterminer si les élèves anglophones utilisent un vocabulaire similaire à celui des élèves français. Puis l'article se focalise sur l'application des logiciels TAL, en définissant les concepts clés de la linguistique de corpus. Les parties quatre et cinq abordent le traitement des données multimodales (écrites et orales) à l'aide d'AntConc à travers des exemples concrets tirés d'une étude de cas. Nous mettons particulièrement en lumière certaines fonctionnalités telles que l'analyse des collocations, des N-grams, l'observation de la fréquence lexicale des apprenants, la distribution d'un mot dans le corpus, etc. L'article se poursuit par une synthèse (partie « Discussion ») autour des apports du corpus pour répondre à la problématique de l'étude, plus précisément concernant la caractérisation de l'interlangue des élèves. Dans la conclusion nous abordons les avantages et les limites des logiciels TAL dans l'analyse des corpus, tout en soulignant leur rôle crucial dans l'évolution des méthodologies en linguistique. Le Glossaire (voir l'annexe 5) donne des précisions terminologiques des notions saillantes et des abréviations de notre étude.

Le dispositif de l'enquête

Le dispositif EMILE en Europe et en France

L'Enseignement de Matières par l'Intégration d'une Langue Étrangère (EMILE) et les sections européennes sont apparus en France dans un contexte d'évolution des politiques éducatives en faveur du bilinguisme et du plurilinguisme. Dès les années 1950, des expérimentations bilingues ont vu le jour pour des raisons géographiques et socio-politiques, notamment le développement des relations économiques sur la frontière franco-allemande (Eurydice, 2019). Ces initiatives se sont amplifiées dans les années 1970, sous l'influence des recommandations de l'Union Européenne.

À partir des années 1990, l'Europe s'est résolument engagée en faveur de l'éducation bi- et plurilingue. Le Conseil de l'Europe a promu des initiatives comme l'EMILE pour renforcer les compétences multilingues des citoyens, avec pour objectif d'accroître la maîtrise de plusieurs langues et de favoriser l'ouverture interculturelle, conformément aux objectifs du Livre Blanc sur l'Éducation et la Formation[3]. Le « multilinguisme » fait référence à la coexistence de différentes langues au sein d'une société, tandis que le « plurilinguisme » désigne la capacité d'un individu à utiliser plusieurs langues (Commission Européenne, 2003, 2005 ; Conseil de l'Europe, 2006). Cette distinction se reflète dans les programmes éducatifs français, qui ont introduit des cours pluridisciplinaires où plusieurs langues sont enseignées dans un même cadre, soutenant une éducation à la fois linguistique et interculturelle.

Comme Duverger (2011, p. 6) l'affirme : « Ces bénéfices ne peuvent être obtenus, ces objectifs ne peuvent être atteints par la simple juxtaposition des enseignements des deux langues et ceux des disciplines : une didactique intégrée, consciente et explicite, s'impose comme une condition indispensable pour la réussite par une intégration […]

entre les didactiques des langues et celles des disciplines ». En effet, la réussite des sections européennes repose sur cette didactique intégrée, où l'enseignement des langues et des disciplines non linguistiques ne sont pas simplement juxtaposés, mais étroitement liés pour maximiser les apprentissages.

Ainsi, ces initiatives, notamment les sections européennes, continuent de se développer dans les collèges et lycées français (voir Annexe 1). Elles jouent un rôle central dans la promotion du plurilinguisme et préparent les élèves aux défis académiques internationaux, tout en leur offrant la possibilité d'obtenir des certifications linguistiques reconnues, comme le Cambridge English Certificate ou le DELF. Cela renforce le lien entre la langue française et les langues étrangères, en plaçant l'intégration didactique au cœur de l'éducation et du développement des compétences interculturelles.

Contexte de l'étude

L'article se situe dans le cadre plus large d'une étude de cas : observation directe d'une classe SELO[4] composée de 19 élèves en Terminale du lycée général et technologique à Chambéry relevant du MEN. 18 heures de cours (l'histoire-géographie enseignée en L2 anglais ; langue de scolarisation L1 est français) ont été observés une fois par mois durant l'année scolaire 2018-2019 en situ. 16 heures de cours ont été enregistrés. Deux professeurs – de langue vivante LV (anglais) – et d'histoire-géographie (DNL) interviennent dans la classe.

Les sections suivantes apportent plus de détails contextuels concernant une mise en place de l'étude de cas et la récolte de données : la présentation du corpus CORINE-SEHEA, des participants, les dates d'observation des cours, les modalités d'enseignement. Nous terminons cette sous-partie par les analyses effectuées en vue d'établir le profil de l'interlangue des apprenants. L'interlangue est le terme introduit pour la première fois par Selinker (1972), dont la définition est donnée ensuite par Quivy et Tardieu-Garnier (2002, p.188) : langue intermédiaire que l'apprenant constitue à partir de tous les matériaux à sa disposition – verbaux ou non verbaux – issus de la langue de départ ou de la langue-cible. Il les intègre dans un système provisoire d'hypothèses sujettes à révision, système qui lui est personnel et manifeste l'état ponctuel de sa connaissance du monde.

Présentation du corpus

Le corpus CORINE a été conçu dans le cadre d'un projet dirigé par John Osborne (professeur émérite de linguistique anglaise au laboratoire LLSETI de l'Université Savoie Mont Blanc). Le corpus CORINE a été réalisé par Evgenia Nicol-Bakaldina à l'Université Savoie-Mont Blanc, Chambéry. Les données ont été recueillies entre septembre 2018 et mai 2019 au lycée général et technologique de Chambéry, relevant du Ministère de l'Éducation Nationale (MEN). Les enregistrements des cours se sont poursuivis jusqu'en juin 2019, et la compilation du corpus s'est déroulée d'avril à décembre 2019, avec des modifications apportées début 2021.

Dans des sections suivantes nous nous arrêtons sur le déroulement du projet (pour plus de détails concernant les étapes de la conception du corpus et des difficultés rencontrées lors de la mise en place du projet, voir voir Nicol-Bakaldina, 2023).

Début du projet

Pour trouver les participants de l'étude de cas, deux options ont été envisagées : contacter les établissements secondaires de Chambéry proposant le dispositif EMILE ou s'appuyer sur des contacts personnels. C'est finalement grâce à un réseau de contacts, développé à partir de l'expérience professionnelle en tant qu'agente contractuelle de l'Éducation Nationale, qu'une professeure d'anglais d'une section européenne à Chambéry a été sollicitée. Après l'accord du proviseur, la préparation du projet a pu commencer (voir Annexe 4 pour la chronologie de l'évolution du projet).

Le premier contact avec les deux professeures (LV et DNL) a eu lieu lors de la visite de l'établissement du 13 septembre 2018. L'objectif de cette rencontre était de présenter le projet de recherche et de discuter des aspects pratiques de l'étude, tels que l'accès aux supports d'enseignement, aux évaluations des élèves et la possibilité d'assister aux cours. Cette première rencontre informelle a également permis d'observer les professeures dans leur environnement de travail, posant ainsi les bases de la collaboration.

Les premières observations des cours EMILE ont eu lieu le 20 septembre 2018. Ces séances « éponges » avaient pour objectif de recueillir des impressions générales sur le déroulement des cours et de commencer à identifier les modalités d'enseignement et les interactions entre professeures et élèves. Cette phase a permis de se familiariser avec le fonctionnement des cours et de poser les premières hypothèses sur l'interlangue des élèves.

Enfin, les modalités de communication avec les professeures ont été établies, principalement via courriel et téléphone, et la fréquence des observations a été planifiée pour permettre une collecte de données régulière sur l'année scolaire.

Observation	Date	Nombre de cours	Discipline observée/commentaires
1-2	20 septembre 2018	2 cours	DNL+LV
3-4	04 octobre 2018	2 cours	DNL+LV
5-6	11 octobre 2018	2 cours	DNL+DNL (prof. LV absente – voyage culturel aux États-Unis)
7-8	15 novembre 2018	2 cours	DNL+LV
9-10	13 décembre 2018	2 cours	DNL+LV
11-12	17 janvier 2019	1 cours	DNL (prof. LV absente – en formation)
13-14	14 février 2019	2 cours	DNL+DNL (évaluation à l'oral en 1-ère heure)
15	14 mars 2019	1 cours	LV (prof. DNL absente – projet dans un autre établissement)
16-17	11 avril 2019	2 cours	DNL+LV
18-19	16 mai 2019	2 cours	DNL+LV
Total		18 cours	11 cours DNL + 7 cours LV

Tableau 1. Dates et nombre des cours observés (observation directe)

Participants

Élèves. Dans le cadre de l'étude de cas, l'échantillon observé est composé d'un groupe de 19 élèves de Terminale inscrits en section européenne. La majorité d'entre eux (75%) n'avait aucune expérience des cours EMILE avant d'intégrer cette section en classe de seconde. Le groupe est constitué d'élèves provenant de quatre classes de Terminale : 12 élèves des trois classes de Terminale S et 7 élèves de la filière ES.

Enseignants. La section européenne du lycée, créée en 2015, est encadrée par deux professeures : l'une enseignant l'anglais, l'autre l'histoire-géographie. La spécificité de cette section réside dans le fait que l'enseignante de langue vivante (LV) est également formée en histoire, ce qui lui permet d'enseigner la DNL (histoire-géographie) en collaboration avec sa collègue. L'enseignante de LV, avec 20 ans d'expérience, joue un rôle de mentor pour sa collègue de DNL. De plus, elle intervient en interdisciplinarité dans d'autres matières au lycée, notamment en physique-chimie en classe de première.

Titre	Professeure de LV	Professeure de DNL
Age	44 ans	49 ans
Diplômes et parcours professionnel	Agrégation d'anglais (1998) ; Double cursus histoire-anglais : - Licence en histoire, Université de la Sorbonne (1995) - DEA, Université de la Sorbonne (1995) ; Classes préparatoires littéraires ; BAC scientifique.	Formation FLE (2017) ; DEA Master (1989) ; Certification supplémentaire en anglais.
Expérience : Nombre d'années d'enseignement de la discipline dans le secondaire	20 ans au lycée X* de Chambéry	18 ans dans secondaire (plusieurs collèges et lycées en France).
Nombre d'années dans la section européenne	4 ans au lycée X de Chambéry, classes de Tle 2017 et 2018	7 ans dans un lycée au sud de la France en 1999-2001 et en 2003-2006 (classes de Seconde, Première, Terminale); 8 ans dans un collège à Chambéry ;
Projets culturels	Projet de l'établissement : voyages aux États-Unis (New York) avec les élèves Tle EMILE en 2017 et 2018	Projet ERASMUS en 1999 : visite de l'Irlande autour du thème « Les femmes en Europe »; projets européens avec les élèves: visites du Parlement Européen à Strasbourg, rencontre avec les eurodéputés à Bruxelles.

Tableau 2. Portrait socio-professionnel des professeures de LV et de DNL

Modalités d'enseignement

L'enseignement de l'histoire-géographie en anglais est assuré en demi-groupes et en classe entière. Chaque élève bénéficie de 3 heures de cours EMILE par semaine (deux fois en classe entière et une fois en demi-groupe), tandis que chaque professeure dispense 2 heures par semaine (une fois avec la classe entière et une fois avec un des deux demi-groupes, voir Tableau 3). Cependant il ne s'agit pas de co-interventions mais d'interventions alternées sur les mêmes sujets ou des sujets enchaînés.

Les deux professeures assurent les cours séparément au sein du même groupe-classe mais elles ont aussi les demi-groupes de cette même classe une fois par semaine Les deux professeures (anglais et histoire-géographie) assurent le même projet pédagogique et les deux enseignent l'histoire-géographie (DNL) au sein de leurs cours EMILE respectifs, donc le niveau de leur collaboration est très élevé.

.

	Professeure LV	Professeure DNL	TOTAL pour deux entretiens
Date de l'entretien	05/12/18	29/10/2018	n/a
Durée	76 min 13 sec	41 min 47 sec	118 min (2 h)
Lieu	Lycée de fonction	Endroit extérieur (café)	Intérieur/Extérieur du lycée
Accord pour enregistrer l'entretien	oui	oui	

Tableau 3. Modalités d'enseignement des cours EMILE en section européenne

Récolte de données

La récolte des données a été une étape fastidieuse. Une partie des supports nous a été transmise directement par la professeure de langue sur une clé USB. D'autres documents ont été récupérés sur des sites anglophones (américains et britanniques) suivant les indications des professeures ou d'après les sources indiquées dans des enregistrements des cours. Ces informations représentent 30 minutes de données en plus de 16 heures d'enregistrement des cours.

Au total 280 unités d'informations ont été récoltées (voir Tableau 4 et Annexe 2) :

- les documents (inputs) utilisés ou crées par les professeures sur un support papier ou projetés au tableau, les traces écrites des professeures au tableau
- les productions écrites et orales (outputs) des élèves : brouillons[5], notes des cours, fiches hybrides[6], évaluations du vocabulaire
- les enregistrements vidéo (observations) proprement dits : enregistrements des cours + auto-enregistrements faits en dehors de la classe

Nombre de supports récoltés : *inputs* + *outputs*	*Inputs* (supports des professeures utilisés en cours)		*Intake* + *Outputs* (productions des élèves)	
	Écrits	Oraux	Écrits	Oraux
	135 : textes, images	6 : audios	95 : les notes prises en cours, les brouillons de préparation des tâches pédagogiques, dont 6 comportant des images (affiches).	11 : audio
	10 : notes au tableau	18 : vidéos	5 : notes au tableau	
Total : 280	169 items		111 items	

Tableau 4. Supports et productions récolés en cours EMILE

Corpus CORINE-SEHEA

Les données ont été transcrites (données orales à l'aide des logiciels TAL Elan et Exmaralda, données écrites – avec AbbyFinerReader11 ou transcrites manuellement), puis le corpus multimodal CORINE-SEHEA[7] a été constitué pour permettre une série d'analyses de l'interlangue des élèves en EMILE. Plus de cinquante mesures explorent l'interlangue des élèves à l'aide des logiciels TAL : Exmaralda, CLAN, AntConc, SketchEngine, UAM Corpus Tool, et Hyperbase. Le travail de recherche étudie dans quelle mesure un statut particulier de la langue en cours immersifs EMILE (à la fois un objet et un outil d'apprentissage) modifie les conditions et les résultats de l'apprentissage. Les analyses ouvrent des pistes de réflexion sur l'enjeu principal d'EMILE : donner une place à la langue étrangère en tant qu'outil d'accès aux connaissances de la DNL[8] pour le plus grand nombre d'élèves possible.

Caractéristiques du corpus. Le corpus, rédigé en anglais, comprend 119 069 mots. C'est l'un des premiers en France à retracer les interactions en classe dans l'enseignement secondaire. Son caractère unique réside dans sa structuration, qui permet d'observer l'environnement linguistique de l'input à l'output (pour les définitions, voir le Glossaire à la fin de l'article). Le corpus est organisé autour des séquences pédagogiques des cours EMILE, qui servent de fil conducteur pour structurer les données. Ces séquences, conçues initialement en 2017 par la professeure de langue vivante (LV) et adaptées en collaboration avec la professeure de discipline non-linguistique (DNL), sont présentées sous forme de textes transcrits et balisés en paragraphes.

Le corpus est actuellement disponible à la demande à l'adresse : eve.nicol83[at]gmail.com.

Les données du corpus sont facilement traçables grâce aux métadonnées : il est possible de voir le nom et la nature du document, sa position dans le corpus (la chronologie générale des cours dispensés), l'ordre du document dans la séquence, la source des données (professeur ou élève), la date du cours, la matière (LV ou DNL), et le nom de la séquence pédagogique dans laquelle le contenu a été enseigné (voir Fig. 1). Le corpus tient également compte du nombre de fois qu'un document a été utilisé en cours[9].

13TOPIC1_Unit1_Séquence2_Doc4BevReportFiveGiants_DNL_DocProf_06sep2018_txt
The Beveridge Report
William Beveridge was part of the war government. The country had a national government made up of all the political parties - the Conservatives, the Labour Party and the Liberals.
The Prime Minister was a Conservative, William Beveridge was a Liberal and there were many Labour Party MPs in high positions in government.
William Beveridge was asked to create the framework for the 'welfare state' , to be implemented at war's end.

Figure 1. Extrait du corpus CORINE avec les métadonnées

Comme le souligne Biber (2010), la représentativité est cruciale dans les études de corpus linguistiques. Le corpus EMILE, sans être exhaustif, est représentatif des cours en section européenne, couvrant un large échantillon d'inputs et outputs d'apprentissage sur une année scolaire. Pour garantir une bonne représentativité, toutes les sources documentaires utilisées en classe européenne ont été incluses.

Limites du corpus CORINE :

- Le corpus ne comprend pas tous les supports pédagogiques utilisés dans les cours EMILE et ne prend pas en compte ce que les élèves ont pu produire au dela des dates d'observation et en dehors de la classe.
- Il n'est pas exclusivement constitué de textes complets, mais inclut parfois des fragments ou des échantillons (comme des extraits de discours, traces écrites des cours).

Malgré ces limites, le corpus reste très représentatif des cours interdisciplinaires en section européenne et offre une base solide pour l'analyse des interactions et des productions linguistiques en classe.

Dans la section suivante nous présentons l'application des logiciels TAL dans la recherche en nous focalisant notamment sur le logiciel AntConc. Ensuite nous illustrons des fonctionnalités les plus saillantes de AntConc en les corroborant avec des exemples tirés du corpus CORINE afin de répondre aux questions de notre étude.

Description des outils utilisés

L’application des logiciels TAL dans la recherche

Dans la recherche de la linguistique du corpus un double type d’exploration de données est possible à l’aide des outils TAL : l’approche montante (corpus-based research) et l’approche descendante (corpus-driven research) (deux termes de Tognini-Bonelli, 2001). En effet, nous essayons d’effectuer une exploration de plusieurs phénomènes, notamment ceux relatifs à l’utilisation de la langue par les élèves en cours CORINE, dans une technique de complémentarité, sorte de « chassé-croisé » :

- L’approche ascendante (corpus-based) est susceptible d’apporter des réponses concernant certains aspects du concept CORINE à partir des questions préalablement posées sur les phénomènes que nous voudrions étudier. Par exemple, nous avons exploré dans quelle mesure l’input des professeurs impacte l’output des élèves, c’est-à-dire, tout ce que les élèves ont retenu et utilisé dans leurs productions – la réception et la restitution des concepts clés en classe européenne.

- L’approche descendante (inductive ou hypothético-déductive, également appelée corpus-driven) permet d’observer les phénomènes récurrents et d’étudier les caractéristiques linguistiques qui émanent du corpus CORINE sans qu’une question précise soit préalablement posée à l’avance. C’est à partir de ces phénomènes observés dans le corpus que nous formulons une série de questions pour étudier un élément souhaité, faisant ainsi du corpus le déclencheur des connaissances. Par exemple, nous avons constaté que l’utilisation du mot government par des élèves varie selon la discipline et le type de production (écrite ou orale). Cela soulève des questions concernant les défis potentiels dans l’appropriation de government, notamment en ce qui concerne les collocations (NN) du mot dans l’input des enseignants et dans les productions des élèves.

Comment explorer le corpus et quelles données est-il possible d’obtenir à l’aide des logiciels TAL ? Les productions sont susceptibles de faire émerger des informations utiles afin de comprendre l’utilisation de la langue par les apprenants. Ces données peuvent renseigner au sujet de la qualité des productions des élèves, des mots et des expressions les plus fréquents, mais aussi permettent d’observer le taux de mots erronés et leur fréquence, les types d’erreurs et leur évolution dans les élocutions des apprenants. Par exemple, nous nous sommes servis du logiciel AntConc dans l’optique de faire émerger les phénomènes les plus fréquents et les plus spécifiques du corpus CORINE (voir aussi les articles portant sur l’extraction et le traitement des données multimodales à l’aide des logiciels CLAN (Nicol-Bakaldina, 2024) et EXMaRALDA (Nicol-Bakaldina, 2023).

L’annotation et le codage des erreurs imposent certaines contraintes dans le calcul du nombre de mots dans le corpus formé par les productions des apprenants. Le calcul s’avère différent en fonction du logiciel utilisé. AntConc compte le nombre des mots, tandis que d’autres logiciels (comme SketchEngine) calcule les lemmes. Par ailleurs, le traitement des erreurs dans le corpus oral devient plus efficace grâce à AntConc, puisque ce logiciel est configurable pour ignorer les erreurs (ou toutes les informations inutiles) mises entre les balises (< >).

Présentation du corpus de comparaison LOCNESS.

Pour établir une comparaison pertinente entre les productions des élèves français et anglophones, le corpus LOCNESS (Louvain Corpus of Native English Essays[10]) a été utilisé comme corpus de référence, ceci afin d’observer si les élèves anglophones se servent du même vocabulaire que les élèves français. Le corpus LOCNESS, composé d'écrits d'élèves britanniques de niveau « A Level » (fin d'études secondaires) et d'étudiants américains et britanniques, totalise 324 304 mots. Afin de garantir une comparaison entre des corpus aux caractéristiques similaires, seuls les écrits d'élèves britanniques de « A Level » portant sur l'histoire, la géographie et la politique ont été sélectionnés, ce qui réduit le corpus de référence à 74 586 mots. Pour but de comparaison, dans le corpus CORINE nous avons sélectionné uniquement des productions écrites des élèves totalisant 9 560 mots.

Même si des corpus de natifs anglophones dans un contexte scolaire sont rares, le corpus LOCNESS est pertinent pour notre étude pour deux raisons. Premièrement, il contient des thèmes spécialisés proches de ceux du corpus CORINE relevant du domaine d’histoire[11]. Deuxièmement, les apprenants des deux corpus ont un âge et un statut social similaires, ce qui renforce la pertinence de cette comparaison.

Le logiciel AntConc

Le logiciel AntConc[12] a été développé par L. Anthony en 2002, depuis cet outil ne cesse pas de gagner de plus en plus le terrain dans la linguistique computationnelle, grâce notamment à son interface intuitive et à son aspect pratique dans la recherche. Si la toute première version 1.0[13] était un concordancier permettant de faire des requêtes de base (générer des lignes de concordance) dans un texte ou dans un groupe de textes, la dernière version (AntConc 4.2.4, sortie en septembre 2023) est complète et sophistiquée. Développé en langage Python et Qte, le programme est compatible avec des systèmes Windows, Linux et MacOS, décrit par l’auteur comme « a multiplatform tool for carrying out corpus linguistics research, introducing corpus methods, and doing data-driven language learning”.[14]

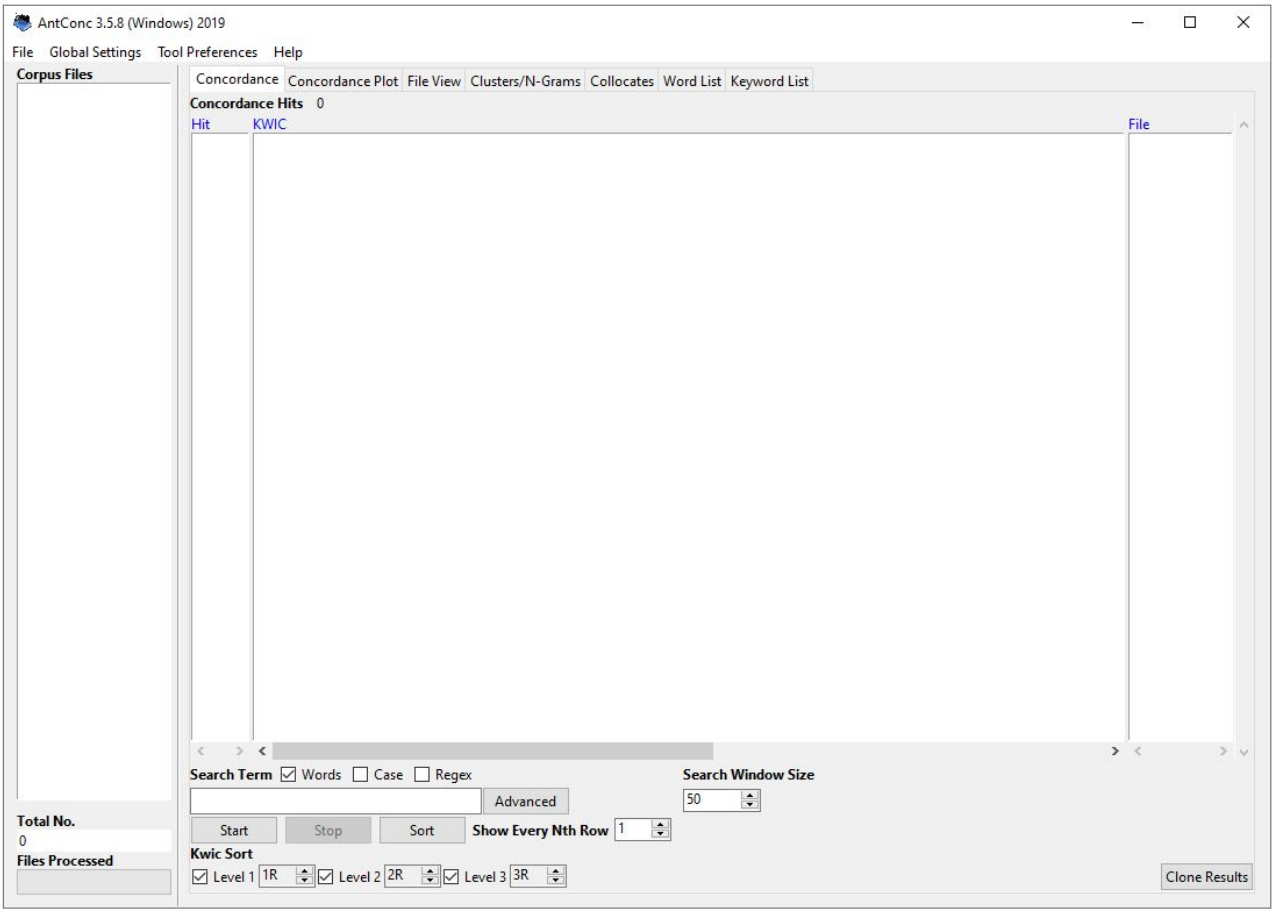


Figure 2. Interface du logiciel AntConc

Application du logiciel AntConc dans la recherche en linguistique de corpus

Le logiciel AntConc, un des plus connus dans le domaine de la linguistique du corpus de par de ces fonctionnalités a été utilisé afin d'explorer le corpus multimodal d'apprenants CORINE. Nous voudrions illustrer sept fonctionnalités de AntConc appliquées au traitement du corpus CORINE en lien avec des questions de notre recherche posées en introduction. Les analyses commencent par des requêtes d'ordre général (mots et expressions fréquents, mots académiques, mots clés) et se focalisent par la suite sur un cas précis en montrant comment le logiciel AntConc peut contribuer à mieux comprendre comment les élèves se servent de la langue et à caractériser leur profil linguistique en cours d'EMILE. La Word List et la Keyword List (corpus de référence LOCNESS) nous permettent d'identifier les mots les plus fréquents et spécifiques, tandis que les collocations et les N-grams révèlent les associations lexicales et les structures récurrentes utilisées par les élèves. La fonction Concordance permet d'examiner le contexte d'utilisation des mots, et la visualisation topologique (Plot) montre leur répartition dans les productions des élèves. Enfin, la fonction Clusters permet d'identifier les groupements de mots fréquents, révélant les structures discursives. Ces

fonctions sont cruciales pour comprendre comment les élèves réutilisent et adaptent le vocabulaire enseigné, tandis que d'autres fonctions d'AntConc (telles que CorpusManager ou StopList) n'étaient pas directement pertinentes pour les questions spécifiques de notre recherche.

Nous commençons par chercher des mots les plus fréquemment utilisés dans le corpus (fonction Word list dans AntConc)

La figure 3 ne montre que les premiers 110 mots du corpus, ce qui permet d'apercevoir la nature des mots employés par l'ensemble de participants. La figure fait émerger des items très variés : des substantifs, des verbes, des pronoms, des adjectifs, des mots grammaticaux, des conjonctions, des chiffres, ce qui reflète des thématiques abordés en cours de la section européenne. Cependant, les caractéristiques langagières diffèrent considérablement selon que le corpus est général ou spécialisé. Il est logique de supposer que le corpus spécialisé dans un domaine précis aurait une tendance à avoir une charge lexicale plus importante dans ce champ concerné qu'un corpus général aux thématiques plus variés.

Rank	Freq	Word	Rank	Freq	Word	Rank	Freq	Word	Rank	Freq	Word	Rank	Freq	Word
1	8525	the	23	538	with	45	373	there	67	228	forty	89	176	many
2	3376	of	24	523	as	46	364	an	68	226	economic	90	176	your
3	3236	to	25	521	have	47	328	countries	69	226	okay	91	172	would
4	2926	and	26	520	war	48	319	all	70	222	when	92	171	good
5	2406	a	27	518	by	49	312	or	71	221	first	93	171	see
6	2288	in	28	499	world	50	309	about	72	221	if	94	171	up
7	1381	is	29	478	from	51	302	do	73	209	also	95	170	nine
8	1352	it	30	463	be	52	295	which	74	209	did	96	170	us
9	1177	that	31	462	two	53	283	who	75	206	will	97	168	crisis
10	1127	you	32	445	i	54	281	has	76	203	uk	98	168	don
11	1046	they	33	439	but	55	271	had	77	200	want	99	167	britain
12	991	was	34	425	can	56	265	she	78	197	british	100	167	like
13	904	s	35	422	x	57	263	five	79	197	why	101	166	right
14	889	nineteen	36	415	were	58	263	yes	80	195	globalisation	102	165	some
15	816	what	37	409	because	59	258	state	81	191	time	103	162	just
16	780	for	38	407	not	60	253	country	82	190	between	104	159	them
17	716	so	39	405	he	61	248	new	83	189	three	105	155	her
18	681	on	40	402	t	62	242	very	84	188	seventy	106	153	thousand
19	650	are	41	389	their	63	240	no	85	184	globalization	107	152	our
20	620	we	42	388	at	64	240	trade	86	183	power	108	152	thatcher
21	613	this	43	381	more	65	234	other	87	178	union	109	151	after
22	539	people	44	377	one	66	233	government	88	177	usa	110	151	united

---

Figure 3. Cent dix mots les plus fréquents du corpus CORINE

Quelles sont les expressions les plus fréquentes employées par les élèves ?

Nous cherchons des séquences des mots/N-grams (séries d'unités : bigram = 2 unités, trigram = 3 unités, N-gram = n unités)[15]. La recherche de N-grams est une fonction intégrée dans la plupart des concordanciers comme AntConc ou SketchEngine. Cependant, d'un logiciel à l'autre il existe des différences sensibles quant au calcul du nombre d'occurrences des N-grams. Ceci peut être expliqué par la façon dont chaque logiciel prend en compte les expressions/combinaisons des mots du corpus. Selon que les combinaisons des mots sont considérées en tant qu'une variation de la même expression ou comme une expression à part (donc un nouveau N-gram) leur nombre peut varier considérablement.

Nous avons utilisé la fonction de N-grams (clusters) de AntConc afin de relever les éléments propres au discours oral dans des productions des apprenants. Figure 4 montre les éléments liés au contenu des cours : de nombreuses dates (in nineteen seventy, nineteen forty-five), les noms des endroits géographiques qui apparaîssent en fonction des sujets étudiés en section européenne ; la description et l'analyse d'une image (we can see) ; les contraintes méthodologiques à respecter : présenter la nature du document (document is a ; is a caricature), structurer le discours (the first document ; the second document).

Par ailleurs, la structure syntaxique du discours peut être tracée grâce aux mots fréquents. Par exemple, le mot that (en tant que pronom relatif ou conjonction de subordination) indique l'utilisation d'énoncés complexes par les élèves (that the Cold [war] ; that globalization is, that globalization can), ou encore la présence des propositions principales connectées avec les conjonctions de coordination and (and we can) et de subordination because, ce dernier introduisant des propositions subordonnées de cause (because of the, because it's).

Rank	Freq	Range	N-gram
1	17	2	nineteen forty five
2	12	2	post war consensus
3	11	2	in nineteen seventy
4	10	2	didn x t
5	8	2	x t want
6	7	2	a lot of
7	7	2	end of the
8	7	2	in nineteen forty
9	7	2	it x s
10	7	2	leader of the
11	7	2	the welfare state
12	6	2	because of the
13	6	2	don x t
14	6	2	nineteen seventy nine
15	6	2	of the country
16	6	2	standard of living
17	6	1	the house of
18	6	1	the nineteen fifties
19	6	2	the standard of
20	5	1	during the war
21	5	1	in the nineteen
22	5	1	nationalisation of industries
23	5	2	nineteen forty seven
24	5	2	of the war
25	5	2	of welfare state
26	5	2	the end of
27	5	2	the labour party
28	5	2	the leader of
29	5	2	the post war
30	5	2	there is a
31	4	2	britain x s

Rank	Freq	Range	N-gram
32	4	2	can x t
33	4	1	education health and
34	4	1	face the future
35	4	2	governing the uk
36	4	1	let x s
37	4	2	nineteen eighty four
38	4	2	nineteen eighty nine
39	4	2	nineteen fifty one
40	4	1	nineteen seventy three
41	4	1	per cent consume
42	4	2	t want to
43	4	2	the cold war
44	4	1	the prime minister
45	3	1	a stalemate along
46	3	1	access to benefits
47	3	1	advantages only for
48	3	1	along the thirty
49	3	1	and infrastructure increase
50	3	1	back the nk
51	3	2	berlin wall nineteen
52	3	1	bring wealth extra
53	3	1	cheap labour leds
54	3	1	costs due to
55	3	1	countries pollution trafficking
56	3	1	doesn x t
57	3	1	due to outsourcing
58	3	1	employment nationalisation of
59	3	2	end of wwii
60	3	1	extra money spent
61	3	1	for education health
62	3	1	for nine months

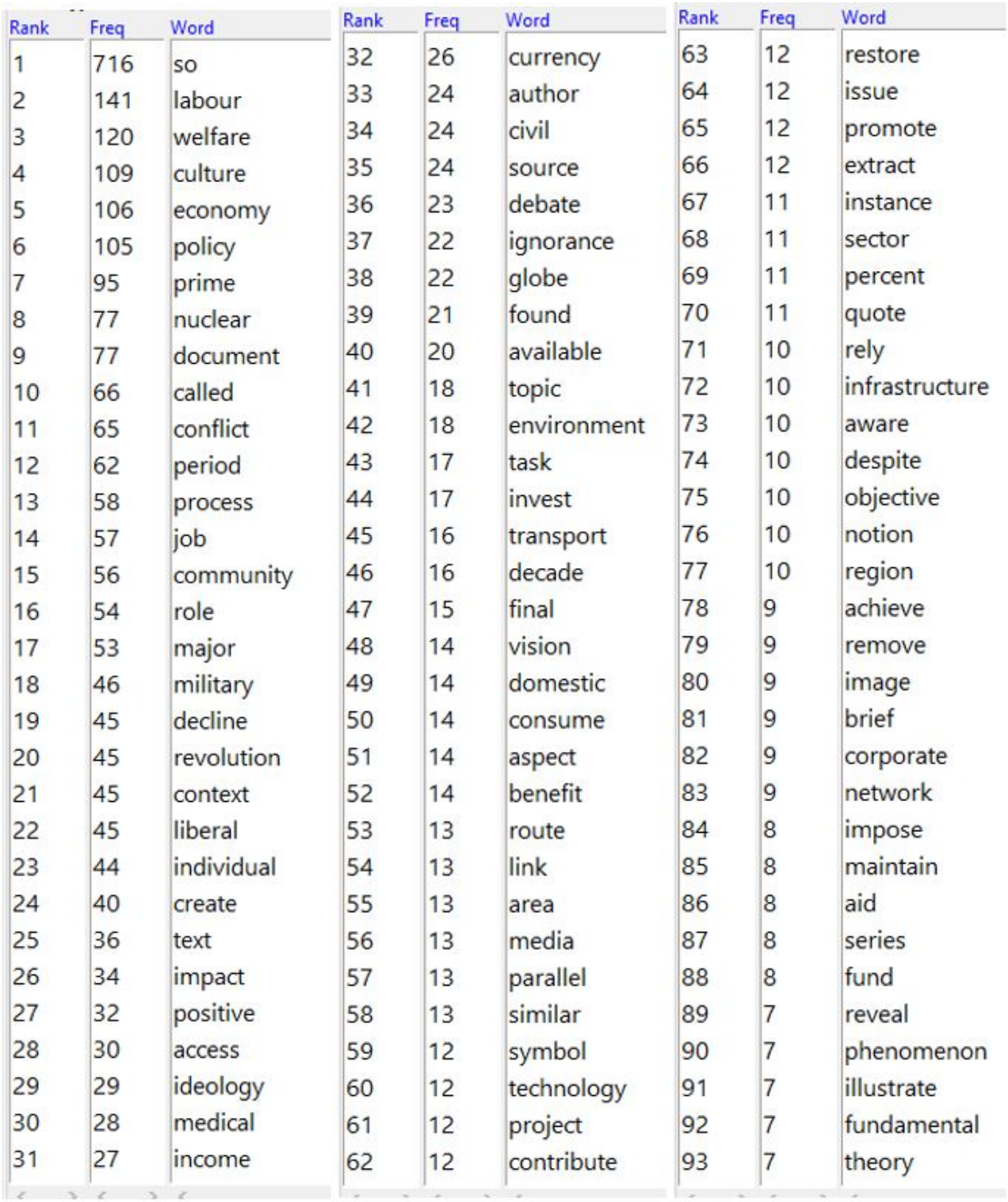
Figure 4. 3-grams les plus fréquents dans les productions des élèves

L’analyse des fréquences lexicales et des N-grams dans le corpus CORINE révèle des éléments clés de l’apprentissage linguistique en contexte EMILE. L’exploration des mots les plus fréquents (Figure 3) montre une grande variété de catégories lexicales, reflétant les thématiques abordées en classe, bien que le corpus, spécialisé en histoire-géographie, affiche une charge lexicale plus ciblée que des corpus plus généraux. D'autre part, l'analyse des N-grams (Figure 4) permet de repérer des expressions récurrentes liées au contenu des cours, telles que des dates, des lieux géographiques, et des structures méthodologiques, tout en mettant en évidence l'utilisation de constructions syntaxiques complexes par les élèves. Ces mots et les expressions les plus fréquemment utilisés sont représentatifs des cours dans le contexte d’EMILE, afin de répondre à une tâche pédagogique demandée. Ces analyses montrent l’impact de

l’input des enseignants sur les productions des élèves et soulignent la manière dont les apprenants s’approprient des éléments linguistiques spécifiques au contexte EMILE.

Chercher des mots académiques, fonction Use specific list below

Afin de déterminer le type du vocabulaire (général, académique ou spécialisé) utilisé par des apprenants dans leurs discours, nous avons comparé le corpus CORINE avec la liste des mots académiques (AWL)[16] utilisée en tant que liste de référence dans AntConc[17] pour voir si les mots académiques les plus fréquents apparaîssent dans notre corpus d’étude.



Rank	Freq	Word
1	716	so
2	141	labour
3	120	welfare
4	109	culture
5	106	economy
6	105	policy
7	95	prime
8	77	nuclear
9	77	document
10	66	called
11	65	conflict
12	62	period
13	58	process
14	57	job
15	56	community
16	54	role
17	53	major
18	46	military
19	45	decline
20	45	revolution
21	45	context
22	45	liberal
23	44	individual
24	40	create
25	36	text
26	34	impact
27	32	positive
28	30	access
29	29	ideology
30	28	medical
31	27	income
32	26	currency
33	24	author
34	24	civil
35	24	source
36	23	debate
37	22	ignorance
38	22	globe
39	21	found
40	20	available
41	18	topic
42	18	environment
43	17	task
44	17	invest
45	16	transport
46	16	decade
47	15	final
48	14	vision
49	14	domestic
50	14	consume
51	14	aspect
52	14	benefit
53	13	route
54	13	link
55	13	area
56	13	media
57	13	parallel
58	13	similar
59	12	symbol
60	12	technology
61	12	project
62	12	contribute
63	12	restore
64	12	issue
65	12	promote
66	12	extract
67	11	instance
68	11	sector
69	11	percent
70	11	quote
71	10	rely
72	10	infrastructure
73	10	aware
74	10	despite
75	10	objective
76	10	notion
77	10	region
78	9	achieve
79	9	remove
80	9	image
81	9	brief
82	9	corporate
83	9	network
84	8	impose
85	8	maintain
86	8	aid
87	8	series
88	8	fund
89	7	reveal
90	7	phenomenon
91	7	illustrate
92	7	fundamental
93	7	theory

Figure 5. Quatre-vingt mots académiques les plus fréquents du corpus CORINE

AntConc a relevé 327 types de mots et 4 001 occurrences de mots académiques (word tokens) dans le corpus CORINE. La figure 5 présente les premiers quatre-vingt d'entre eux (ayant la fréquence la plus importante). Les résultats affichés mettent en avant plusieurs choses. Premièrement, une grande majorité des mots académiques sont des substantifs du champ sémantique socio-economique : labour, policy, culture, welfare,

economy, community, job, currency, income, decline, infrastructure, etc. Ensuite, le champ sémantique à connotation politique et militaire est aussi représenté par des substantifs et des adjectifs comme : nuclear, conflict, military, revolution, ideology. Les verbes académiques les plus fréquents sont : created, consume, contribute, remove, invest. Les mots comme process, debate et access nécessitent une vérification au cas par cas, car ils peuvent comprendre plusieurs parties de discours (les verbes et les substantifs) en même temps.

Outre ces mots qui relèvent du contenu des cours, nous trouvons aussi :

- les mots propres à la didactique, comme debate, document[18] :
- des consignes comprises dans les titres des supports : Reading comprehension, oral debate ;
- des consignes des professeures à l'oral : Make a debate with your neigbour ; You just look at the document and try to answer the question ; And now focus on the document entitled…;
- la méthodologie de l'analyse des supports appliquée par des élèves lors des productions : The second document describes a man who works…

Globalement nous constatons que les mots académiques dont la fréquence est importante sont très peu nombreux. Au-delà de l'adjectif major à la dix-septième position dans le corpus la fréquence de toutes les unités descend en dessous de cinquante occurrences. Ceci laisse penser que le nombre de mots académiques du corpus CORINE est relativement bas, ce qui indique à son tour que le vocabulaire des cours observés est probablement hautement spécifique.

Chercher des mots clés, fonction key words

Les mots clés qui émergent dans le corpus CORINE par rapport au corpus de référence LOCNESS (Fig.6) renvoient aux contextes particuliers de leurs emplois. Plus particulièrement, nous pouvons accéder au champ didactique et tracer les thèmes spécifiques abordés en cours, comme The Welfare State, Trade Unions and miners' strikes, The nationalisation of industries, Political party, Economic crisis, wages and employment, etc. Dans un sens plus large ce lexique est le reflet des programmes d'enseignement propres aux sections européennes dans le secondaire en France. Par ailleurs, la présence très faible des mots académiques laisse suggérer que la place la plus importante est occupée par un vocabulaire plus ciblé, hypothèse qui sera confirmée par des analyses supplémentaires du corpus.

Rank	Freq	Keyness	Effect	Keyword
1	121	+ 420.3	0.0236	nineteen
4	34	+ 106.3	0.0067	forty
6	26	+ 92.52	0.0051	usa
7	25	+ 88.96	0.0049	labour
8	38	+ 79.85	0.0075	five
9	19	+ 67.6	0.0037	uk
10	22	+ 65.25	0.0043	seventy
11	23	+ 64.35	0.0045	welfare
12	18	+ 64.04	0.0035	cent
13	30	+ 63.14	0.0059	want
14	17	+ 60.48	0.0033	globalisation
15	16	+ 56.92	0.0031	access
16	16	+ 56.92	0.0031	nationalisation
17	16	+ 56.92	0.0031	sixty
18	25	+ 52.99	0.0049	trade
19	18	+ 51.78	0.0035	fifty
20	22	+ 51.4	0.0043	employment
21	14	+ 49.81	0.0028	crisis
22	29	+ 48.88	0.0057	she
23	29	+ 47.29	0.0057	money
24	25	+ 46.77	0.0049	increase
25	15	+ 46.25	0.003	churchill

Rank	Freq	Keyness	Effect	Keyword
26	13	+ 46.25	0.0026	consensus
27	15	+ 46.25	0.003	eighty
28	20	+ 45.29	0.0039	nine
29	22	+ 44.25	0.0043	full
30	15	+ 41.79	0.003	miners
31	22	+ 40.33	0.0043	economy
32	25	+ 40.02	0.0049	you
33	13	+ 39.41	0.0026	coal
34	11	+ 39.13	0.0022	exchange
35	11	+ 39.13	0.0022	mps
36	11	+ 39.13	0.0022	nhs
37	11	+ 39.13	0.0022	ninety
38	69	+ 36.79	0.0133	war
39	10	+ 35.57	0.002	futur
40	22	+ 33.86	0.0043	free
41	43	+ 33.84	0.0084	state
42	16	+ 33.36	0.0031	per
43	16	+ 33.36	0.0031	post
44	11	+ 32.62	0.0022	inflation
45	11	+ 32.62	0.0022	ussr
46	14	+ 32.21	0.0028	private
47	14	+ 32.21	0.0028	twenty
48	9	+ 32.01	0.0018	nuclear
49	9	+ 32.01	0.0018	wwii
50	11	+ 28.71	0.0022	reduce

Figure 6. Cinquante mots clés relevés dans le corpus écrit des élèves CORINE (corpus de référence LOCNESS)[19]

En conclusion, les analyses des fréquences lexicales, des N-grams, des mots académiques et des mots clés dans le corpus CORINE révèlent des spécificités importantes du vocabulaire utilisé en contexte EMILE, ainsi que des caractéristiques marquantes de l'interlangue des élèves.

Tout d'abord, l'analyse des mots les plus fréquents et des N-grams met en lumière un vocabulaire varié, incluant à la fois des mots grammaticaux et des termes propres aux thématiques abordées en cours, tels que des noms de lieux géographiques ou des dates historiques. Ces expressions récurrentes, souvent liées aux contenus pédagogiques, démontrent l'appropriation du lexique enseigné en classe, bien que la structure

syntaxique reste relativement complexe chez les élèves.

L'analyse des mots académiques montre une présence limitée de ce type de vocabulaire dans le corpus, suggérant que les cours EMILE observés utilisent un lexique plus spécifique et contextualisé, orienté autour des thématiques socio-économiques et politiques abordées en section européenne. Les substantifs liés à ces champs (ex. policy, labour, economy) prédominent, tandis que les verbes académiques sont rares, ce qui souligne une focalisation sur des concepts et termes spécialisés.

Enfin, les mots clés extraits en comparaison avec le corpus LOCNESS reflètent le contenu didactique et les thèmes abordés en section européenne, confirmant le caractère spécifique et spécialisé du lexique enseigné dans ces classes. Ces analyses montrent que l'interlangue des élèves est marquée par un vocabulaire thématique et ciblé, mais demeure en développement sur le plan de la richesse lexicale académique, ce qui reflète à la fois les apports de l'input des enseignants et les limitations des productions des apprenants.

Recherches spécifiques. Cas particulier du mot « government »

Dans cette sous-partie, nous souhaitons montrer comment l'utilisation du logiciel AntConc permet de suivre l'appropriation du mot government des inputs (enseignants) aux outputs (productions d'élèves). Bien que government ne soit pas le mot le plus fréquent du corpus CORINE (avec 233 occurrences, 66ème position dans la liste), il se classe parmi les six substantifs les plus couramment employés par les deux enseignants de LV et de DNL ET repris par les élèves dans leurs productions : people (fréquence 343), war (f.323), state (f.182), government (f.169) et UK (f.154) (la fréquence indique le nombre de fois que le mot a été utilisé par les deux professeurs ensemble dans leurs discours). Ainsi le mot government constitue un excellent point d'entrée pour étudier plus en profondeur l'acquisition et l'utilisation du vocabulaire en contexte EMILE allant de l'input vers l'output.

Chercher un mot dans un texte

Cette série des requêtes s'avèrent très utiles pour répondre aux questions de notre recherche en terme d'acquisition et d'utilisation du lexique en cours EMILE : la recherche d'un mot du corpus (ici : government) utilisé tout seul et/ou en collocation avec d'autres mots, sa forme (nom, gérondif, aspect verbal, etc.), chercher un contexte de son énonciation donné à l'origine par les professeurs, puis tracer des modifications éventuelles dans l'emploi de ce mot par des élèves.

Compléter des requêtes simples afin d'effectuer des recherches plus avancées en utilisant la fonction wordcard/joker :

- le joker « * » est utile pour la recherche des formes de verbes (p.ex. gov* = governs, governed);

- le joker « + » permet de chercher zéro ou un signe après le joker (p.ex. government+ = government agreed);

- le joker « @ » fait des recherches qui portent sur un mot entier : (p.ex. the @ government = the British government, the new government...)

Utiliser les expressions régulières :

- chercher tous les mots qui commencent par une séquence souhaitée : p.ex. bgov = government

- chercher tous les mots qui se terminent par une séquence souhaitée : p.ex. entb = government, contentment, document

- chercher des mots qui contiennent des séquences souhaitées, p.ex. tous les mots qui commencent par « n » et accessoirement qui contiennent « a » et/ou « t », et/ou « s » (p.ex. bn[ats]= nationality, names, name, nature, nationhood)

- le motif [ ] va chercher (tous, ou un des) caractères qui se trouvent entre les crochets, p.ex. [cha])

Utiliser des mots ciblés pour des recherches avancées dans un texte avec la fonction Use search term from list below, ou encore Use context words and horizons

(cliquer sur le champ Advanced , saisir le mot recherché, cliquer sur Apply pour valider le choix; AntConc va chercher les mots indiqués cinq caractères à gauche et/ou à droite à côté des mots choisis dans la première fonction)

Dans quelle mesure les mots communément employés par les professeures dans leurs cours (LV et DNL) sont-ils répétés dans les productions des élèves ? La fonctionnalité « plot » de AntConc abordée ci-dessous donne des éléments de réponse.

Chercher des substantifs fréquents repris par des apprenants dans leurs productions, fonctionnalité plot[20] (« visualisation topologique »)

AntConc affiche les résultats de recherche en concordance sous forme de « code-barres », avec la longueur du texte (calculée en chars) normalisée par rapport à la largeur de la barre, chaque occurrence (hit) est représentée par une ligne verticale à l'intérieur de la barre. La fonction plot permet de visualiser la position des résultats de recherche dans les textes cibles.

La fonction nous permet d'observer l'appropriation lexicale par les élèves en cours interdisciplinaires. Par exemple, nous avons constaté que l'utilisation du mot spécifique government par des élèves varie selon la discipline et le type de production (écrite ou orale). Cela soulève des questions (cf.corpus-driven research) concernant les défis potentiels dans l'appropriation de government, notamment en ce qui concerne les collocations (NN) du mot dans l'input des enseignants et dans les productions des élèves.

Nous avons observé si les élèves utilisent plus souvent dans leurs productions le lexique donné par la professeure de langue ou par la professeure d'histoire-géographie. L'analyse est effectuée en traçant l'origine du mot grâce à la fonction de AntConc « visualisation topologique » (plot) ainsi qu'en observant les collocations du mot (fonction Collocations) et des tâches pédagogiques dans les cours où les deux enseignantes aborde le même thème. Cela permet de voir si le contexte d'utilisation du mot change de l'input à l'output, ce qui contribuera à notre réflexion sur les conditions spécifiques de l'apprentissage et sur la complexité du dispositif EMILE.

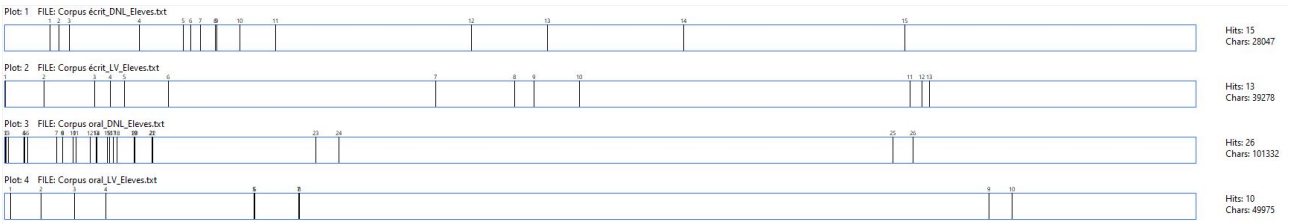


Figure 7. Distribution du mot fréquent government dans le corpus des élèves

Le mot fréquent government figure dans l'output des élèves est communément employé par les deux professeures. Comment ce mot est-il employé dans les productions des apprenants ? Le substantif est employé 64 fois dans les productions des élèves (Fig. 7). D'après la figure 7 ce substantif est le plus souvent utilisé par des élèves à l'oral en cours d'histoire-géographie (26 occurrences), suivi dans l'ordre décroissant de fréquence par le sous-corpus écrit, également en cours de DNL (15 occurrences), puis en cours de langue à l'écrit (13 occurrences) et à l'oral (10 occurrences).

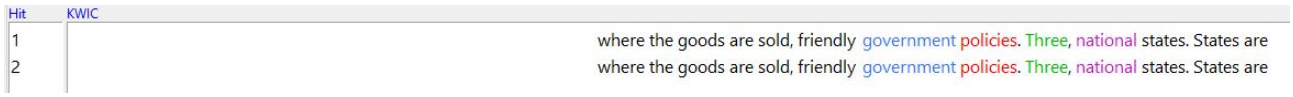
Même si les élèves ont recours au mot government communément utilisé dans les deux disciplines, sa fréquence totale en DNL reste supérieure après relativisation : 25,6 occurrences dans les productions en DNL versus 16,7 en LV pour 10 000 mots (soit 21 % d'écart entre deux disciplines)[21]. Par ailleurs, l'utilisation de ce mot est presque deux fois plus fréquente à l'écrit que à l'oral quelle que soit la discipline. Une analyse plus approfondie montre que le substantif government est repris par des élèves dans les deux disciplines mais l'output se fait dans un degré variable selon la discipline et le type de production dans les sous-corpus : écrit ou oral. La question se pose, par conséquent, par rapport à un autre défi possible dans l'appropriation de government : les

collocations du mot (NN) dans l'input des professeures et dans les productions des élèves.

Chercher des collocations avec government (NN), fonction Collocations

Nous allons puiser plus profondément dans l'analyse de l'appropriation du mot government en classe EMILE. En anglais il est souvent employé dans des compositions nominales NN (government policy, etc.). Des modèles de cet usage sont-ils présents dans l'input, le modèle est-il repris par les élèves ? Nous avons observé quatre types de collocations dans les inputs donnés par les professeures : government intervention, government polic* (policy/policies), inter-government organisation, government ministers.

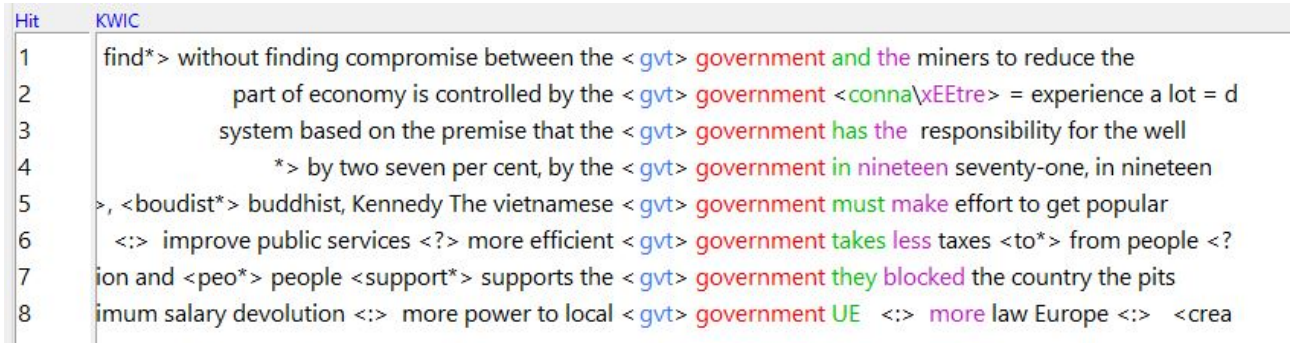
La Figure 8 illustre une collocation de type NN (« government policies ») dans une production d'élève en cours de DNL, issue de l'exercice sur la mondialisation et les entreprises transnationales (« Globalisation and TNCs »). Cette production met en évidence la réutilisation du lexique enseigné, mais aussi la manière dont les élèves modifient ou adaptent les structures linguistiques apprises à partir des discours des professeurs. Dans les productions (outputs) des élèves, une seule collocation autour de government (NN) a été trouvée : government policies (dans l'énoncé : "There are many factors attracting TNCs to a country, cheap raw materials, cheap labour supply, good transport, access to markets where the goods are sold, friendly government policies"). Cet exemple montre concrètement comment les apprenants se servent des inputs reçus et comment ces collocations (NN) apparaissent dans leurs productions. En ce sens, l'exemple permet de souligner le défi d'appropriation du lexique enseigné, en particulier lorsqu'il s'agit de collocations nominales courantes, et de montrer l'impact des cours interdisciplinaires EMILE sur la production linguistique des élèves.



Hit	KWIC
1	where the goods are sold, friendly government policies. Three, national states. States are
2	where the goods are sold, friendly government policies. Three, national states. States are

Figure 8. Collocation de type NN dans les productions des élèves

Pour compléter cette analyse, nous avons également étudié l'usage de l'abréviation gvt (abréviation de government) dans les productions écrites des élèves (Fig. 9). Dans leurs brouillons, les élèves abrègent souvent des substantifs pour noter rapidement l'essentiel. Pendant la transcription, nous avons remplacé ces abréviations par le mot complet government afin de standardiser les données, permettant ainsi de les analyser avec AntConc.



Hit	KWIC
1	find*> without finding compromise between the < gvt> government and the miners to reduce the
2	part of economy is controlled by the < gvt> government <conna\xEEtre> = experience a lot = d
3	system based on the premise that the < gvt> government has the responsibility for the well
4	*> by two seven per cent, by the < gvt> government in nineteen seventy-one, in nineteen
5	>, <boudist*> buddhist, Kennedy The vietnamese < gvt> government must make effort to get popular
6	<:> improve public services <?> more efficient < gvt> government takes less taxes <to*> from people <?
7	ion and <peo*> people <support*> supports the < gvt> government they blocked the country the pits
8	imum salary devolution <:> more power to local < gvt> government UE <:> more law Europe <:> <crea

Fig. 9. Collocations avec « gvt » dans des productions écrites des élèves

Nous avons obtenu ds GN (groups nominaux) suivants: efficient government, local government, the Vietnamese government ce qui met en évidence l'appropriation partielle du lexique enseigné dans le cadre des cours EMILE. Cette observation révèle que, bien que les élèves abrègent souvent le mot government dans leurs brouillons, ils continuent de reproduire des structures nominales complexes et pertinentes, démontrant ainsi une certaine maîtrise du lexique spécialisé dans un contexte historique et géopolitique. Ceci suggère la réutilisation de plus en plus fréquente et diversifiée des collocations nominales enseignées, bien que cette appropriation demeure partielle et parfois modifiée en fonction du contexte.

Discussion

Les analyses menées sur le corpus CORINE à l'aide du logiciel AntConc apportent des réponses précises aux questions de recherche sur l'interlangue des élèves en contexte EMILE. Plusieurs aspects ont été explorés pour comprendre les processus d'apprentissage linguistique, en particulier la nature du lexique utilisé, sa fréquence et distribution selon les disciplines (LV et DNL), et la manière dont les élèves réutilisent les mots enseignés dans des contextes variés.

Appropriation du lexique et interlangue des élèves

- Le vocabulaire utilisé dans les cours EMILE est caractérisé comme majoritairement spécialisé (analyse validée par l'extraction des mots clés avec le corpus de référence des natifs anglophones LOCNESS) et lié aux thématiques abordées, telles que l'histoire, la géographie, l'économie et la politique (constat fait par d'autres chercheurs, voir par exemple Tran Tat, 2022)

- L'input des enseignants joue un rôle crucial dans l'acquisition du vocabulaire spécialisé. Parmi les onze substantifs communément utilisés par les deux enseignants, dix apparaissent dans les productions des élèves, Par exemple, des termes com.me government (233 occurrences), state (182 occurrences), policy et

economy sont couramment utilisés, reflétant des concepts centraux dans l'apprentissage des élèves. Sur les 43 substantifs employés par les enseignants[22], 27 (soit 63 %) se retrouvent dans les productions des élèves, ce qui démontre la place centrale que le vocabulaire occupe dans les cours EMILE en soulignant l'impact de l'input cumulatif sur l'apprentissage.

- L'exemple du mot government montre que les élèves réutilisent le vocabulaire dans des contextes similaires à ceux des enseignants, mais aussi dans des contextes plus créatifs et adaptés aux exigences des tâches scolaires. Par exemple, des collocations comme government policies sont réemployées dans les productions écrites des élèves, bien que des groupes nominaux telles que efficient government ou local government apparaissent également, témoignant d'une appropriation progressive et flexible. Cette utilisation diversifiée montre que les élèves commencent à intégrer le vocabulaire enseigné tout en l'adaptant aux besoins pratiques, comme la prise de notes rapide, où des abréviations comme gvt, couramment employées à l'écrit.

Stratégies d'apprentissage

Les professeures recourent à de multiples stratégies d'enseignement afin de faciliter l'appropriation du lexique :

1. Transfert des aspects pragmatiques du langage. Les professeures en EMILE parviennent à stimuler la prise de parole des élèves malgré les moyens limités d'expression en L2 (gestion de l'approximation linguistique, terme de Gajo) grâce notamment à l'interaction dont le taux est très élevé dans les deux disciplines observées (LV et DNL). Des productions orales de « mise en situation » (compétence pragmatique) s'organisent souvent en forme de dialogues. Des tâches finales et intermédiaires de production orale mettent l'élève dans un contexte de simulation proche de la réalité en lien avec le sujet étudié.

2. Transfert des stratégies métacognitives et métalinguistiques, notamment en apprentissage du vocabulaire (mémorisations, exercices à trous, repérages des mots clés) et de l'analyse des supports (capacité de structurer et d'organiser des idées : «Compare and contrast the documents»). Ces stratégies – jeux de rôle, simulations, missions à accomplir – permettent d'approcher les situations de la vie réelle évoquées par Coyle: « successful language learning can be achieved when people have the opportunity to receive instruction, and at the same time experience real-life situations in which they can acquire the langage more naturalistically » (Coyle, 2010, p.11).

Ces observations révèlent que le processus d'apprentissage en EMILE est caractérisé

par une appropriation partielle mais croissante du vocabulaire spécialisé et des collocations enseignées. Les élèves réutilisent fréquemment le lexique des enseignants, en particulier dans les disciplines DNL, tout en montrant des signes d'adaptation et de créativité, notamment dans les productions écrites. Cette appropriation est renforcée par les interactions fréquentes entre enseignants et élèves, ainsi que par les stratégies pédagogiques utilisées, comme les jeux de rôle et les simulations, qui favorisent un apprentissage actif du vocabulaire.

En somme, l'interlangue des élèves en cours EMILE se caractérise par une réutilisation du vocabulaire spécifique enseigné, mais aussi par des adaptations créatives en forme de nouveau lexical blocks (terme de Biber, 2006) qui témoignent de leur progression dans l'appropriation des structures lexicales complexes. Les expressions et les mots utilisés sont indissociables de la thématique abordée, et de ce fait très représentatifs des cours immersifs EMILE. Par ailleurs, l'input cumulatif des enseignants et l'output des élèves sont étroitement liés, formant un cycle d'apprentissage où les productions des élèves deviennent elles-mêmes un input pour leurs pairs.

Conclusion

Les analyses effectuées avec AntConc montrent à la fois les avantages et les limites de cet outil dans l'exploration d'un corpus multimodal CORINE en linguistique appliquée. AntConc se distingue par ses fonctionnalités efficaces pour analyser des éléments linguistiques comme les fréquences lexicales, les collocations et les concordances, permettant de mieux comprendre l'interlangue des élèves en contexte EMILE (enseignement de matières par intégration d'une langue étrangère). Cet outil facilite le suivi des items lexicaux et du vocabulaire spécialisé fréquemment utilisés dans des contextes thématiques spécifiques, tels que l'histoire et la géographie, offrant des éclairages précieux sur l'acquisition lexicale des élèves dans ces domaines. Enfin, AntConc est très utile lors de traitement des erreurs dans le corpus, car l'information que l'on ne souhaite pas afficher (par ex. les mots non-normés, marqueurs d'hésitations, etc.) peut être mise entre les balises, ainsi ces informations ne seront pas prises en compte par AntConc lors d'un calcul des mots ou de génération des lignes de concordance.

Cependant, AntConc présente aussi des limites dans la gestion de données multimodales complexes, car il ne prend en charge que des formats texte basiques (TXT, HTML ou XML), nécessitant ainsi des étapes de prétraitement supplémentaires pour rendre compatibles divers types de données. Le fait de doubler les entrées textuelles - mettre les informations non-souhaitées (ex.erreurs) entre les balises toute en proposant la version corrigée juste après - influe sur la taille du corpus et peut éventuellement fausser la comptabilisation des données du corpus traité par d'autres logiciels.

Malgré ces défis, AntConc reste une ressource précieuse pour la recherche en linguistique appliquée, soutenant le développement de méthodologies affinées pour les études de corpus sur l'interlangue. Ses fonctions analytiques directes fournissent des

données essentielles pour comprendre l'adaptation lexicale et la créativité linguistique dans l'acquisition de langues en contexte disciplinaire, répondant ainsi aux besoins croissants de la linguistique numérique et des études de corpus. Nous espérons par notre travail participer activement à une réflexion nécessaire concernant l'utilisation du corpus formé par les productions des apprenants dans le but d'étudier le fonctionnement de la langue, ainsi que pour la formation possible des professeurs qui travaillent avec la textométrie. Le champ de recherche en lien avec notre corpus d'étude reste ouvert, pour contribuer à une réflexion sur les usages des corpus de linguistique en France, mais aussi sur l'avenir du dispositif EMILE, tout en prenant en compte l'avantage essentiel qu'il offre – la langue comme un moyen d'accès aux connaissances de la DNL.

Bibliographie

Anthony, L. (2021). AntConc (Version 4.0.0) [Computer Software]. Tokyo, Japan: Waseda University. URL: https://www.laurenceanthony.net/software.html

Benabbes, S. (sans date). Du brouillon au texte final : accompagner les élèves du secondaire pour une amélioration de la pratique scripturale en FLE. Les Cahiers de Didactique des Lettres [En ligne]

Biber, D. (2006). University language: A corpus – based study of spoken and written registers. John Benjamins.

Blom, J.P., et Gumperz, J.J. (1972). Social meaning in linguistic structures: code-switching in Norway. Dans : J.J. Gumperz, et D.H. Hymes (dir.), *Directions in Sociolinguistics*. (p. 407–434). Blackwell.

Conseil de l'Europe (2006). Cadre européen commun de référence. Apprendre, enseigner, évaluer. Les Éditions Didier.

Corder, S. P. (1981). Error analysis and interlanguage. Oxford University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-238.

Coyle, D., Hood, P., et Marsh, D. (2010). *Content and Language Integrated Learning*. Cambridge University Press.

Dostie, G. et Lefeuvre, F. (dir.) (2017). Lexique, grammaire, discours. Les marqueurs discursifs. Honoré Champion.

Duverger, J. (dir.). (2011). Enseignement bilingue. Le professeur de « Discipline Non Linguistique ». Statut, fonctions, pratiques pédagogiques. ADEB - Association pour le Développement de l'Enseignement Bi/plurilingue. URL : https://blogacabdx.ac--bordeaux.fr/lvr64/wp-content/uploads/sites/86/2019/07/ Enseignement-bilingue-Le--professeur-de-discipline-non-linguistiqueADEB.pdf

European Commission. (2003). *Promoting language learning and linguistic diversity : Action plan 2004-06*. Publications Office of the European Union. URL:http://op.europa.eu/en/publication-detail/-/publication/b3225824-b016--42fa-83f6-43d9fd2ac96d

European Commission. (2005). *Communication from the Commission to the Council, the European Parliament, The European economic and Social Committee and the Committee of Regions. A New Framework Strategy for Multilingualism*. URL : https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2005:0596:FIN:en:PDF

Eurydice. (2019). *Chiffres clés de l'enseignement des langues à l'école en Europe : édition 2017*. Education, Audiovisual and Culture Executive Agency. URL: https://data.europa.eu/doi/10.2797/082121

Gajo, L. (2007). Linguistic Knowledge and Subject Knowledge: How Does Bilingualism Contribute to Subject Development? *The International Journal of Bilingual Education and Bilingualism*, *10*(5), 563-581.

Galatanu, O. (2021). Les marqueurs illocutionnaires holophrastiques du désaccord : sémantisme et polyphonie fonctionnelle d'une classe de phraséologismes pragmatiques. Lexique, 29, 75-95.

Granger, S., Gilquin, G., et Meunier, F. (dir.) (2015). The Cambridge handbook of learner corpus research (1er ed.). Cambridge University Press.

Johns, T. (1991). Should you be persuaded. Two samples of data-driven learning materials. English Language Research Journal. 4, 1-16.

Marsh, D. (2012). *Content and Language Integrated Learning (CLIL). A development trajectory* [thèse de doctorat]. Université de Cordoue.

Nicol-Bakaldina, E. (2024). Logiciel CLAN : transcription et traitement des données multimodales. *Mélanges Crapel ATILF-CNRS 44*(2), 153-175

URL :https://www.atilf.fr/wp-content/uploads/publications/-MelangesCrapel/Melanges_44_2_7_Nicol-Bakaldina.pdf

Nicol-Bakaldina, E. (2023 a). L'enseignement d'une matière par intégration d'une langue étrangère (E.M.I.L.E) en France : le rôle et l'utilisation de la langue à l'intersection entre deux disciplines dans l'enseignement secondaire [thèse de doctorat, Université Savoie Mont Blanc, LLSETI]. URL : https://hal.science/tel-04252531

Nicol-Bakaldina, E. (2023 b). Logiciel EXMARaLDA : outil de traitement des données discursives orales. Mélanges Crapel ATILF-CNRS 44(1), 330-348. URL :https://www.atilf.fr/wpcontent/uploads/publications/MelangesCrapel/-Melanges_44_1_17_Bakaldina-Nicol_2023.pdf

Rastier, F. (2004). Enjeux épistémologiques de la linguistique de corpus. *Texto !, 9*(2). URL : http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html

Rohlfing, K., Loehr, D., Duncan, S., Brown, A., Franklin, A., Kimbara, I., Milde, J.-T., Parrill, F., Rose, T., Schmidt, T., Sloetjes, H., Thies, A. et Wellinghoff, S. (2006). Comparison of multimodal annotation tools. Gersprachsforschung, 7, 99-123.

Selinker, L. (1972). Interlanguage. International Review of Applied Linguistics, 10(3), 219-231.

Steuckardt, A. (2018). Les marqueurs de reformulation formés sur dire : exploration outillée. Langages, 212, 17-34.

Tognini-Bonelli, E. (2001). Corpus linguistics at work. John Benjamins Publishing.

Tran Tat, N. (2022). Intégration d'une langue cible (cas de l'anglais) et appropriation des savoirs disciplinaires en physique-chimie : co-construction d'une séquence autour de l'effet de serre et du réchauffement climatique en terminale scientifique [Thèse de doctorat, Université Paris Cité]. HAL. URL : https://theses.hal.science/tel-03929908/

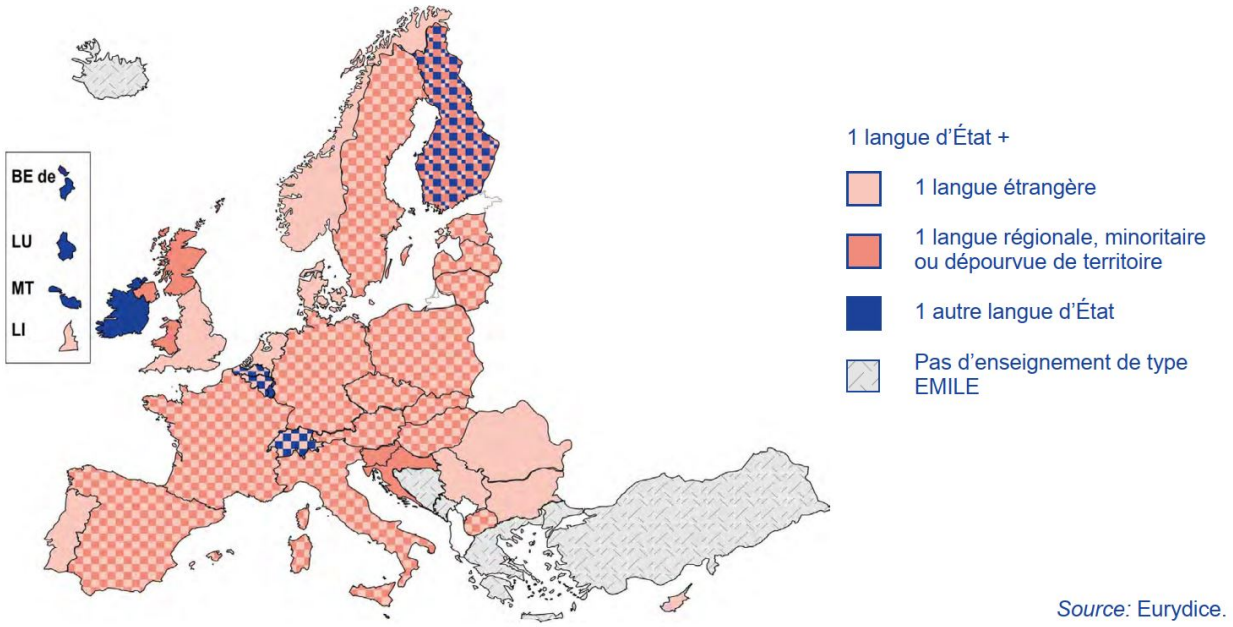
Tutin, A., Jaques, M-P., Kraif, O. et Hartwell, L. (sans date). Introduction à la linguistique de corpus [Cours en ligne]. Université Grenoble Alpes. FUN (France Université Numérique).

URL : https://www.fun-mooc.fr/fr/cours/introduction-a-la-linguistique-de-corpus/.

Quivy, M., et Garnier-Tardieu, C. (2002). Glossaire de didactique de l'anglais (2e éd.). Ellipses.

Annexes

Annexe 1. Statut des langues cibles enseignés dans le cadre de l'EMILE dans l'enseignement primaire et/ou secondaire général, 2015/2016 (Eurydice)



Source: Eurydice.

Annexe 2. Sources d'informations initiales, type et quantité de données obtenues

	Source d'informations initiales	Type de données recueilli	Données obtenues (output des élèves)
Échantillon (19 élèves)	enregistrements vidéos	-trace écrite au tableau ; -discours des élèves en classe : cours et évaluations (réponses, débats, discussions, monologues)	productions écrites : 95 Productions orales : 16 heures de cours + 11 fichiers d'auto-enregistrement ;
	photographies	-notes de cours (trace écrites) faites en classe ; -brouillons de préparation des tâches pédagogiques ; -fiches de travail « hybride » ; -évaluations écrites (fiches de vocabulaire ; affiches).	
	enregistrements sonores	-évaluations orales (auto-enregistrements hors classe)	

Annexe 3. Fiche des enregistrements des cours EMILE

Nom	Date	Type	Taille	Durée
CoursDNL_04oct_H...	06/10/2018 19:14	MP4 Video File	396 864 Ko	00:24:43
CoursDNL_04oct_H...	04/10/2018 12:04	AIMP: MPEG-4 Au...	11 636 Ko	00:12:16
CoursDNL_11avr_H...	12/05/2020 18:07	MPEG Video File	168 314 Ko	00:23:22
CoursDNL_11avr_H...	12/05/2020 17:57	MPEG Video File	112 536 Ko	00:21:35
CoursDNL_11oct_H...	13/04/2020 22:06	MPEG Video File	204 500 Ko	00:23:23
CoursDNL_11oct_H...	13/04/2020 22:24	MPEG Video File	186 860 Ko	00:22:56
CoursDNL_11oct_H...	13/04/2020 22:28	MPEG Video File	211 406 Ko	00:23:22
CoursDNL_11oct_H...	13/04/2020 22:31	MPEG Video File	151 204 Ko	00:19:20
CoursDNL_13decH2...	18/04/2020 15:09	MPEG Video File	168 858 Ko	00:23:22
CoursDNL_13decH2...	18/04/2020 15:11	MPEG Video File	104 678 Ko	00:14:14
CoursDnl_14fev_H1...	27/04/2020 14:03	MPEG Video File	148 962 Ko	00:23:22
CoursDnl_14fev_H1...	27/04/2020 13:58	MPEG Video File	49 032 Ko	00:06:36
CoursDnl_14fev_H2...	27/04/2020 13:47	MPEG Video File	186 688 Ko	00:23:22
CoursDnl_14fev_H2...	27/04/2020 13:51	MPEG Video File	131 980 Ko	00:17:08
CoursDnl_14fev_H2...	27/04/2020 13:59	MPEG Video File	4 358 Ko	00:00:32
CoursDNL_15nov_H...	17/04/2020 09:44	MPEG Video File	195 336 Ko	00:23:23
CoursDNL_15nov_H...	17/04/2020 09:47	MPEG Video File	149 920 Ko	00:16:43
CoursDNL_16mai_H...	21/05/2020 22:03	MPEG Video File	154 666 Ko	00:23:22
CoursDNL_16mai_H...	21/05/2020 22:08	MPEG Video File	133 184 Ko	00:23:22
CoursDNL_17jan_H...	20/04/2020 16:58	MPEG Video File	161 982 Ko	00:22:25
CoursDNL_17jan_H...	20/04/2020 17:02	MPEG Video File	183 738 Ko	00:23:22
CoursDNL_17jan_H...	20/04/2020 16:54	MPEG Video File	180 108 Ko	00:23:22
CoursDNL14fev_H1...	27/04/2020 14:15	MPEG Video File	160 410 Ko	00:23:22
CoursLV_04oct_H1P1	04/07/2019 20:00	MP4 Video File	254 719 Ko	00:29:50
CoursLV_04oct_H1P2	05/07/2019 12:33	MP4 Video File	196 234 Ko	00:23:05
CoursLV_11avr_H1P1	12/05/2020 17:37	MPEG Video File	159 570 Ko	00:23:23
CoursLV_11avr_H1P2	12/05/2020 18:25	MPEG Video File	155 544 Ko	00:23:22
CoursLV_11avr_H1P3	12/05/2020 18:26	MPEG Video File	16 840 Ko	00:04:02
CoursLV_13dec_H1P1	18/04/2020 15:16	MPEG Video File	170 038 Ko	00:23:23
CoursLV_13dec_H1P2	18/04/2020 16:08	MPEG Video File	148 516 Ko	00:23:22
CoursLV_13dec_H1P3	18/04/2020 16:08	MPEG Video File	15 560 Ko	00:01:36
CoursLV_15nov_H1P1	17/04/2020 09:51	MPEG Video File	175 764 Ko	00:23:22
CoursLV_15nov_H1P2	17/04/2020 10:33	MPEG Video File	143 978 Ko	00:19:53
CoursLV_16mai_H1P1	21/05/2020 21:45	MPEG Video File	144 020 Ko	00:23:10
CoursLV_16mai_H1P2	21/05/2020 21:50	MPEG Video File	142 532 Ko	00:23:22
CoursLV_17mars_H...	02/04/2022 21:15	MP4 Video File	620 361 Ko	00:23:22
CoursLV_17mars_H...	03/04/2022 09:30	MP4 Video File	512 572 Ko	00:23:22
CoursLV_17mars_H...	03/04/2022 09:32	MP4 Video File	16 682 Ko	00:00:45

Annexe 4. Chronologie de l'évolution du projet

ÉTAPE	Période/date	Détails
Phase théorique et méthodologique	septembre 2018 septembre 2018 sep 2018 – jan 2019 juillet-septembre 2018 septembre 2018 septembre 2018	- élaboration du projet de recherche ; - choix de la méthodologie de recherche ; - étude de la littérature (théorique et méthodologique) au sujet d'EMILE ; - recherche de l'établissement ; - accord du proviseur/des professeurs ; - demande d'une autorisation parentale d'enregistrement de l'image/de la voix des élèves.
Préparation et mise en place du projet	13 sep 2018 septembre 2018 oct – nov 2018 sep 2018 - déc 2018	- premier contact avec les enseignants sur place ; - établissement des dates/séances d'observation ; - conception des outils de récolte des données : questionnaires ; interviews ; - choix des outils de traitement et d'analyse des données.
Phase pratique : recueil des données, observation proprement dite	septembre 2018 – mai 2019 décembre 2018 mars 2019 mars et mai 2019	- observation des cours (cf. détails des séances observées ci-dessous) ; - réalisations des interviews : avec professeur de DNL ; avec professeur de LV ; - administration des questionnaires ; - récolte des productions écrites ; - évaluations.
Validation de la méthodologie de recherche	jan 2019 – nov 2019	- validation interne, externe, du construit.
Phase de traitement et d'analyse des données obtenues	jan 2019 – nov 2019 jan 2019 - sep 2020	- traitement des données (transcription des matériaux, constitution du corpus de linguistique) ; - analyses des données obtenues (corpus, documents, interviews, questionnaires, enregistrements).
Phase de présentation, de validation, d'interprétation et de discussion des résultats	jan 2021 – juin 2021 2021	- présentation des résultats des analyses (des documents, des corpus, des interviews, des questionnaires) ; - validation des résultats ; - évaluation de l'efficacité du dispositif EMILE; - discussion des thèmes saillants du projet.
Phase de conclusion. Finalisation du travail de recherche	jan 2021 – juin 2021	- présentation du bilan du travail et réponses aux questions/problématique de la recherche ; - validation des hypothèses émises en amont ; - perspectives et limites du projet.
Rédaction du rapport (thèse) Mise en forme et relecture	sep 2020 - juin 2022 sep 2022 – mai 2023	Rédaction de différentes parties (chapitres) du projet

Annexe 5 : Glossaire

AWL (Academic Word List) : Liste des mots académiques les plus fréquemment utilisés dans des corpus académiques, utilisée comme référence pour évaluer le niveau de spécialisation du vocabulaire des apprenants (Coxhead , 2002).

Champ didactique : Domaine d'étude qui s'intéresse aux méthodes et pratiques d'enseignement, notamment comment une langue ou une discipline est enseignée et apprise.

Collocations : Combinaisons de mots qui apparaissent fréquemment ensemble dans un corpus, révélant des associations linguistiques naturelles.

Concordance : Fonctionnalité d'un logiciel TAL (Traitement Automatique de la Langue) qui permet de visualiser les occurrences d'un mot ou d'une expression dans un corpus et de les analyser dans leur contexte.

Corpus : un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications. […] Le corpus se résume à un échantillon de la langue, un réservoir d'exemples ou d'attestations (Rastier, 2004).

Corpus-based research : Approche ascendante dans laquelle des questions précises sont posées à partir d'un corpus afin de vérifier des hypothèses prédéfinies.

Corpus-driven research : Approche descendante qui consiste à observer les phénomènes linguistiques récurrents dans un corpus sans hypothèses préalables, permettant de générer de nouvelles questions de recherche.

Discipline linguistique : Ensemble des méthodes et approches utilisées pour l'étude de la langue et du langage, comme la syntaxe, la sémantique, la morphologie, etc.

DNL (Discipline Non-Linguistique) : Réfère aux matières enseignées en langue étrangère dans le cadre de l'EMILE, comme l'histoire-géographie, qui ne sont pas des cours de langue mais qui sont enseignées dans une langue autre que la langue maternelle des élèves.

Input : Données linguistiques entrantes : écrite ou orale (dispensée par un professeur, un élève, un locuteur natif).

Intake : Données linguistiques acquises par l'élève.

Items lexicaux : Unités de vocabulaire (mots, expressions) que les apprenants doivent acquérir. Ils peuvent inclure des substantifs, des verbes, des adjectifs, ou des

expressions spécifiques à une discipline.

LV (Langue Vivante) : Terme désignant les langues étrangères enseignées dans le cadre du système scolaire. Dans le contexte EMILE, la LV correspond à la langue utilisée pour enseigner une discipline non-linguistique.

Mots cles : Mots qui sont particulièrement fréquents dans un corpus étudié par rapport à un corpus de référence, ce qui indique leur importance dans le contexte étudié.

Lemmatisation : Processus qui consiste à regrouper les formes fléchies d'un mot sous sa forme de base ou radicale, facilitant l'analyse linguistique.

N-grams : Séquences de N unités (mots, phrases) dans un corpus qui révèlent des structures linguistiques récurrentes. Un bigram est composé de deux mots, un trigram de trois mots, et ainsi de suite.

NN (Noun-Noun Collocation) : Terme utilisé en linguistique pour désigner des collocations où deux noms (substantifs) sont associés dans une expression. Par exemple : government policy.

Output : Production de l'élève par suite des étapes de l'input et de l'intake. « Output is a realization of productive language skills » (Blom et Gumperz, 1972 ; Marsh, 2012, p.152).

Notes

1 https://www.universalis.fr/encyclopedie/traitement-automatique-des-langues/

2 Des exemples sur l'utilisation d'autres outils TAL ainsi que des éléments plus précis développés dans l'article peuvent être consultés dans nos travaux : Nicol-Bakaldina 2023(a), Nicol-Bakaldina 2023(b), Nicol-Bakaldina 2024

3 https://op.europa.eu/en/publication-detail/-/publication/d0a8aa7a-5311-4eee-904c-98fa541108d8/language-fr

4 Section Européenne et Langues Orientales. A noter que l'appellation SELO est officiellement abandonnée en faveur de celle de l'EMILE à partir de la rentrée 2019 suite à la reforme du lycée de 2018.

5 Brouillon (production écrite d'un élève) est considéré comme « un écrit intermédiaire, entre un premier écrit et le suivant, entre un élève et les autres, entre un oral, une lecture et un nouvel écrit, entre une expérience et sa mise à distance » (Benabbes, S., s.d.).

6 Les fiches de travail conçus par des professeurs expressément pour des cours EMILE et qui contiennent des espaces dédiés à la prise de notes, aux réflexions personnelles, aux réponses des élèves. Ce sont des traces écrites (textuelles) constituées par des professeures au tableau lors des activités pédagogiques : des mots clés, explicitation d'une notion difficile, question à laquelle les élèves doivent répondre, consignes...

7 CORpus des INteractions à l'Ecole : Section Européenne Histoire Enseignée en Anglais. Pour la simplicité d'usage, nous avons conservé dans cette note de recherche la première partie uniquement : CORINE.

8 La Discipline Non Linguistique

9 Par exemple lorsqu'un support est exploité par les professeures de deux matières, il va figurer deux fois dans le corpus intégral. De même, l'exploitation d'un support pendant le cours a l'incidence sur le nombre des entrées textuelles dans le corpus (ex. la lecture à une haute voix par un des élèves est l'output pour celui qui parle et l'input pour ceux qui écoutent).

10 https://www.learnercorpusassociation.org/resources/tools/locness-corpus/

11 ex. le système parlementaire, la monarchie britannique, la guerre en Iraq, la seconde guerre mondiale, etc.

12 https://www.laurenceanthony.net/software/antconc/

13 https://www.laurenceanthony.net/software/antconc/releases/AntConc100/help.pdf

14 https://www.laurenceanthony.net/software/antconc/releases/AntConc424/help.pdf

15 La création d'une liste des N-grams les plus fréquents du corpus permet d'identifier un phénomène linguistique qui aurait pu sinon rester invisible. Dans les recherches plus spécialisées les N-grams peuvent aider à observer les marqueurs discursifs ou les phraséologismes (expressions fixes propres au contexte étudié).

16 Academic Word List; liste de mots académiques, établie par Coxhead (2002).

17 Démarche: Tool Preferences → Word list → Use specific list below.

18 Le substantif image après la vérification est écarté puisqu'il relève du contenu des

cours, aucun emploi de ce mot pour des raisons méthodologiques n’est observé.

19 Voir la présentation du corpus LOCNESS dans la section consacrée à la description des outils mobilisés. Pour cette analyse, uniquement les outputs (productions écrites) des élèves ont été sélectionnées. Pour plus de détails sur la comparaison avec le corpus des natifs anglophones LOCNESS, voir notre thèse (Nicol-Bakaldina, 2023)

20 https://antconc-manual.readthedocs.io/en/latest/concordance_plot.html

21 Calcul : 8,9 ÷ (25,6 +16,7) x 100

22 Analyse effectué avec AntConc puis trié manuellement