Actes des 11èmes Journées Internationales de la Linguistique de Corpus

3-7 juillet 2023

Grenoble, Franc

Comités

Comité scientifique

Tatiana ALEKSANDROVA (Université Grenoble Alpes, Lidilem)

Sara ALVAREZ-MARTINEZ (Université Grenoble Alpes, ILCEA4)

Georges ANTONIADIS (Université Grenoble Alpes, Lidilem)

Alex BOULTON (Université de Lorraine, ATILF)

Shirley CARTER-THOMAS (Université Sorbonne Nouvelle-Paris 3, LaTTiCe)

Cristelle CAVALLA (Université Sorbonne Nouvelle, DILTEC)

Florence CHENU (CNRS, DDL)

Franck CINATO (CNRS, HTL)

Cosimo De GIOVANNI (Università degli Studi di Cagliari, Italie)

Corinne DENOYELLE (Université Grenoble Alpes,LITT&Arts)

Anne DISTER (Université Saint-Louis, Bruxelles, Belgique)

Sascha DIWERSY (Université Paul Valéry, Praxiling)

Gaétane DOSTIE (Université de Sherbrooke, Canada)

Patrick DROUIN (Université de Monréal, OLST)

Céline DUGUA (Université d'Orléans, LLL)

Iris ESHKOL-TARAVELLA (Université Paris Nanterre, MoDyCo)

Emmanuelle ESPERANCA-RODIER (Université Grenoble Alpes, LIG)

Carole ETIENNE (CNRS, ICAR)

Cécile FREROT (Université Grenoble Alpes, Lidilem)

Laure GARDELLE (Université Grenoble Alpes, Lidilem)

Francis GROSSMANN (Université Grenoble Alpes, Lidilem)

Céline GUILLOT-BARBANCE (ENS de Lyon, IHRIM)

Laura HARTWELL (Université Toulouse Capitole, LAIRDIL)

Lydia-Mai HO-DAC (Université Toulouse Jean Jaurès, CLLE)

Agata JACKIEWICZ (Université Paul Valéry, Praxiling)

Marie-Paule JACQUES (Université Grenoble Alpes, Lidilem)

Olivier KRAIF (Université Grenoble Alpes, Lidilem)

Frédéric LANDRAGIN (CNRS, LaTTiCe)

Jean-Philippe MAGUÉ (ENS de Lyon, ICAR)

Pascale MOUNIER (Université Grenoble Alpes, LITT&Arts)

Hilary NESI (Coventry University)

Samia OUNOUGHI (Université Grenoble Alpes, Lidilem)

Christophe PARISSE (Université Paris Nanterre, MODYCO)

Claude PONTON (Université Grenoble Alpes, Lidilem)

Céline POUDAT (Université Côte d'Azur, BCL)

Sophie PRÉVOST (CNRS, LaTTiCe)

Elisa RAVAZZOLO (Università di Trento)

Josette REBEYROLLE (Université Toulouse Jean Jaurès, CLLE)

Solange ROSSATO (Université Grenoble Alpes, LIG)

Isabelle ROUSSET (Université Grenoble Alpes, Lidilem)

Yan RUI (Université Grenoble Alpes, Lidilem)

Julie SORBA (Université Grenoble Alpes, Lidilem)

Agnès STEUCKARDT (PRAXILING)

Agnès TUTIN (Université Grenoble Alpes, Lidilem)

Denis VIGIER (ICAR)

Geoffrey WILLIAMS (Université Grenoble Alpes, Litt&Arts)

Virginie ZAMPA (Université Grenoble Alpes, Lidilem)

Comité d’organisation

Présidente

Marie-Paule Jacques (Université Grenoble Alpes, LIDILEM)

Membres

Sara Alvarez Martinez (Université Grenoble Alpes, Ilcea4)

Florence Chenu (CNRS, DDL)

Sascha Diwersy (Université Paul Valéry, Praxiling)

Emmanuelle Esperanca-Rodier (Université Grenoble Alpes, LIG)

Carole Etienne (CNRS, ICAR)

Cécile Fabre (Université De Toulouse Jean Jaurès, Clle)

Cécile Frerot (Université Grenoble Alpes, Lidilem)

Lucia Gomez (Université Grenoble Alpes, Lidilem)

Sylvain Hatier(Université Grenoble Alpes, Lidilem)

Marie-Paule Jacques (Université Grenoble Alpes, Lidilem)

Olivier Kraif (Université Grenoble Alpes, Lidilem)

Thomas Lebarbe (Université Grenoble Alpes, Litt&Arts)

Jean-Philippe Magué (ENS De Lyon, ICAR)

Samia Ounoughi (Université Grenoble Alpes, Lidilem)

Claude Ponton (Université Grenoble Alpes, Lidilem)

Solange Rossato (Université Grenoble Alpes, Lig)

Isabelle Rousset (Université Grenoble Alpes, Lidilem)

Julie Sorba (Université Grenoble Alpes, Lidilem)

Agnès Tutin (Université Grenoble Alpes, Lidilem)

Geoffrey Williams (Université Grenoble Alpes, Litt&Arts)

Rui Yan (Université Grenoble Alpes, Lidilem)

Virginie Zampa (Université Grenoble Alpes, Lidilem)

Plénières

Linguistique de corpus et pragmatique linguistique : opportunités et difficultés

Jérôme Jacquin

Maître d'enseignement et de recherche, Section des Sciences du Langage et de l'Information, Université de Lausanne

La linguistique de corpus a longtemps été associée à l'analyse quantitative de données textuelles. Aujourd'hui, on admet plus volontiers non seulement la réalité et l'intérêt des méthodes mixtes articulant démarches quantitatives et qualitatives, mais aussi la diversité des matériaux langagiers sur lesquels appliquer ces techniques d'exploration. Alors que la première limite a été dépassée par une réflexion épistémologique plus générale sur les objectifs de la linguistique de corpus (par ex. Egbert, Larsson, et Biber 2020; Mayaffre 2005; Rastier 2004), la seconde l'a plutôt été au travers de l'émergence, l'agrégation et la (encore très relative) mutualisation de corpus oraux voire multimodaux (par ex. Adolphs et Carter 2013; Avanzi, Béguelin, et Diémoz 2016; Baldauf-Quilliatre et al. 2016).

Cela dit, la dimension pragmatique reste encore largement sous-explorée en linguistique de corpus, du moins comparativement aux niveaux plus traditionnels que sont les niveaux phonologiques, morphosyntaxiques, transphrastiques et textuels, et cela malgré quelques exceptions notables (par ex. Aijmer et Rühlemann 2014; Romero-Trillo 2008; Rühlemann 2019). Cela est possiblement dû à la diversité et à la nature des observables en jeu ainsi qu'à l'hétérogénéité théorique et méthodologique qui caractérisent le champ de la pragmatique. Certains courants influents, comme l'analyse conversationnelle d'inspiration ethnométhodologique, sont par ailleurs fermement réfractaires à la quantification (Schegloff 1993).

Plus fondamentalement, la difficulté d'une « pragmatique de corpus » (Aijmer et Rühlemann 2014) réside probablement dans la nécessité d'une définition et d'une délimitation claires non seulement des unités explorées, mais aussi du contexte (ou du niveau de granularité du contexte) pertinent de ces unités. On peut citer trois enjeux particulièrement complexes : (i) ce qui relève des implicatures et donc des contenus implicites et procéduraux (par ex. « cesser » > [avoir commencé] ; ou encore les connecteurs et autres marqueurs discursifs) ; (ii) ce qui relève de la séquentialité et des problèmes de récursivité qu'elle pose (le contexte d'une unité comme un déictique ou un acte de langage devient l'unité d'un contexte plus large, par exemple transphrastique, textuel ou interactionnel) ; (iii) ce qui relève de la multimodalité et de sa contribution à la production et à l'interprétation de conduites sémiotiques plus larges, dans la mesure où les comportements paraverbaux questionnent la discrétisation et la description d'unités bien moins structurées, voire grammaticalisées, que les conduites verbales.

De manière à exemplifier ces enjeux théoriques et méthodologiques relatifs à l'application de la linguistique de corpus à des observables pragmatiques, la conférence se basera pour partie sur un projet de recherche en cours financé par le Fonds National Suisse

(www.unil.ch/sli/posepi; [100012_188924]). D'une durée de quatre ans, le projet entend proposer une exploration quantitative et qualitative des marqueurs épistémiques et évidentiels du français-en-interaction. Le corpus étudié rassemble 28h d'interactions naturelles en français vidéo-enregistrées en Suisse romande et documentant des débats politiques (9h de débats publics et 5h de débats télévisés) et des réunions professionnelles (14h de réunions de coordination et de brainstorming dans des entreprises de communication, architecture et ingénierie). Les données ont été intégralement transcrites et révisées dans le logiciel ELAN (Wittenburg et al. 2006) et en mobilisant les conventions de transcription ICOR (Groupe ICOR 2013). La conférence sera l'occasion de revenir sur les aspects les plus pertinents du guide d'annotation élaboré pour cette recherche (Jacquin et al. 2022a) et de problématiser la présente réflexion à partir de deux études de cas (Jacquin 2022; Jacquin et al. 2022b).

Les grands corpus oraux d'interactions : où, quand et comment les faire intervenir en Didactique Des Langues étrangères ?

Florence Mourlhon-Dallies

Professeure en Sciences du langage et Didactique des langues, EDA (Education, Discours, Apprentissages) et GRIP (Global Research Institute of Paris), Université Paris Cité

La didactique des langues (DDL) se définit volontiers comme une discipline d'emprunt. Parmi les « disciplines mères », il est courant de convoquer la linguistique. Or ces dernières années, la linguistique s'est engagée dans la constitution et l'étude de grands corpus écrits et oraux. Si certains corpus littéraires et médiatiques (pour la presse écrite) ont été intégrés aux préoccupations didactiques, on constate à l'inverse que les grands corpus oraux peinent à entrer dans les manuels et n'ont qu'un faible impact sur les pratiques d'enseignement. Pourtant, on peut penser que les corpus oraux, dont les corpus d'interactions sur lesquels nous mettons le focus, sont précieux quand on veut assurer la transposition didactique des savoirs linguistiques en objets d'enseignement - tant pour les futurs formateurs et enseignants de langue que pour les apprenants étrangers plus directement.

A partir de notre propre expérience mais aussi de ressources construites à des fins didactiques par des linguistes et des didacticiens (CLAPI Fle, CLAPI Corail, CLAPI Interfare, FLEURON, FLORAL-PFC) nous aborderons les questions suivantes : Auprès de quels publics utiliser les grands corpus oraux ? A quelles fins ? De quelles manières ? Puis, pour finir, nous interrogerons quelques notions clés de la DDL qui paraissent fortement impactées par le recours aux grands corpus oraux, dont celle de l'authenticité (versus du réalisme) des documents d'appui et celle de l'exposition discursive des apprenants « à l'oral ».

Communications Orales & Posters

Sommaire

Constitution d'un corpus de français parlé en Tunisie pour l'étude des marqueurs discursifs

Mariem ABID
Unité de recherche Clesthia EA 7345, Université Sorbonne Nouvelle
meriem.abid@sorbonne-nouvelle.fr

Mots Clés : marqueur discursif, corpus oral, transcription, annotation, Tunisie, francophonie

1 Introduction

Nous nous proposons dans cette recherche d'étudier les marqueurs discursifs (désormais MDs) en français parlé en Tunisie. Étant donné que ces petits mots apparaissent essentiellement à l'oral, nous avons décidé de constituer un corpus du français oral tunisien à partir d'enregistrements d'entretiens, comme le remarquent Dostie et Pusch : « Les MD sont des mots particulièrement usités dans la langue orale » (Dostie & Pusch, 2007 : 4). Ils sont « monnaie courante de l'oral » (Dostie, 2004) et sont susceptibles d'être plus utilisés à l'oral qu'à l'écrit. Notre choix s'est porté sur l'oral principalement pour cette raison qui nous a paru logique. Constituer notre propre corpus de conversations authentiques semble être la solution la plus adéquate pour l'étude des MDs. Il faut rappeler que la Tunisie ne dispose pas d'un corpus de français parlé. En ce qui nous concerne et dans le cadre de notre thèse, nous allons exploiter ce corpus pour l'étude des MDs employés par les locuteurs tunisiens. Cependant, le corpus pourrait être exploitable aussi pour étudier d'autres phénomènes sur les différents niveaux linguistiques que ce soit syntaxique, lexical, phonétique, pragmatique... La constitution de ce corpus nous permettra de comparer les MDs utilisés en français parlé en France et ceux utilisés en français tunisien.

Il est important de rappeler que les recherches sur les MDs ont été fructueuses durant ces dernières années (Dostie & Pusch, 2007: 3). Néanmoins, elles ont essentiellement porté sur le français parlé par des locuteurs natifs, notamment le français hexagonal et le français canadien (D. Vincent, 1993 ; Dostie, 2004 ; Lefeuvre 2012, 2017, 2020). À notre connaissance, aucun chercheur ne s'est penché sur ce phénomène dans un corpus parlé par des locuteurs tunisiens.

2 Corpus et méthodologie

Dans cette partie, nous nous focalisons sur les démarches et les méthodes adoptées dans la construction, la transcription et l'annotation du corpus.

2.1 Corpus

Étudier les MDs en français parlé nécessite de pouvoir s'appuyer sur un corpus oral. Or, à l'heure actuelle, aucun corpus de français parlé en Tunisie n'est disponible. Nous avons donc décidé de constituer notre propre corpus dans le cadre de notre thèse de doctorat.

Pour la collecte des données langagières orales, nous nous sommes inspirée de la méthodologie de la constitution du corpus de français parisien (CFPP2000.) et sur le guide des

bonnes pratiques de Baude et al. (Baude et al., 2006). Notre objectif est d'avoir le maximum de contenu langagier, notre choix s'est donc naturellement porté sur les entretiens semi-directifs qui nécessitent la présence du chercheur dont le rôle est de solliciter le participant pour le faire parler. Un questionnaire a été prévu pour l'occasion. Mais il ne s'agit pas de le suivre à la lettre. Le profil du participant est déterminant dans la formulation des questions. Actuellement, la taille de notre corpus est de 13 heures. Il s'agit d'un corpus en cours de construction car l'objectif final est de 20 heures. Concernant la durée moyenne de l'entretien, elle est de 25 minutes. Mais cela peut varier d'un entretien à un autre, certains ont duré 15 minutes alors que d'autres ont duré 1 heure. Ce choix de diversification de durée est dû à notre volonté d'interroger le plus de locuteurs et en même temps d'avoir des entretiens plus ou moins longs. Cela peut dépendre aussi du profil de l'interviewé et de ses capacités communicationnelles. Actuellement, 32 personnes ont participé aux enregistrements : 16 hommes et 16 femmes âgés de 21 à 67 ans.

2.2 Méthodologie

La première étape de la constitution du corpus est de chercher et de sélectionner des participants qui correspondent à nos critères de choix des interviewés. Nous avons établi deux critères principaux qui reposent principalement sur l'approche variationniste de notre étude. Premièrement, le participant doit parler français couramment. Son niveau doit correspondre à un niveau B1/ B2 selon CECRL (Cadre Européen Commun de Référence pour les langues). Il doit être locuteur natif de l'arabe dialectal tunisien et vivant en Tunisie pour s'assurer que son français n'a pas été influencé par celui d'un autre pays, particulièrement le français hexagonal. Ensuite, la deuxième étape consiste à choisir le type et la thématique d'échange. Étant donné que la langue française n'est pas une langue de communication courante en Tunisie, le choix de la conversation spontanée nous a paru peu efficace. Il n'existe pas de situation de communication qui fait intervenir des Tunisiens en train de parler français entre eux d'une manière spontanée. La présence d'un interviewer (le chercheur) s'avère être obligatoire dans la mesure où il sera là pour guider l'entretien et pousser les interviewés à parler. L'entretien directif nous semble ne pas être pertinent puisque qu'il ne donnera pas une totale liberté aux participants. Nous avons alors choisi le juste milieu qui est l'entretien semi-directif de type dialogal. La thématique sera la même pour tous les entretiens par souci de comparabilité. Le matériau doit être homogène et comparable. 4 thématiques principales ont été abordées lors des entretiens : vie personnelle et professionnelle, vie quotidienne, politique et réseaux sociaux. Celles-ci sont communes à la majorité des entretiens, mais d'autres sujets sont également évoqués selon le profil et l'intérêt du participant. En ce qui concerne les métadonnées, elles ont été générées automatiquement grâce à un formulaire en ligne envoyé et rempli par les interviewés. C'est ce qui nous permettra de prendre en considération les informations extralinguistiques au moment de l'analyse telles que la région d'origine, l'âge, le sexe, le rapport à la langue française, etc.

3 Résultats

Les enregistrements sonores recueillis représentent des données brutes primaires non exploitables d'un point de vue scientifique. Leur transformation en corpus doit passer par des étapes nécessaires. Il est indispensable de passer par une transcription et une annotation qui vont permettre l'utilisation du corpus (Abouda & Baude, 2006). Pour ce faire, nous avons commencé par transcrire nos données en suivant le protocole de transcription des ESLOs (Guide Transcripteur, v4 : 2013) avec des adaptations. Nous avons effectué une transcription

orthographique à l'aide du logiciel Transcriber. Une première annotation consistant à segmenter le discours a été faite au moment de la transcription. Il faut savoir que Transcriber permet de faire 3 types de découpage. Le premier découpage consiste à diviser chaque entretien en section correspondant aux différentes thématiques. La plupart des entretiens ont été découpés en 4 parties qui correspondent aux 4 thèmes choisis (exemple : section 1 : vie personnelle ; section 2 : vie professionnelle…). Le second découpage correspond à une segmentation en tour de paroles. À chaque changement de locuteur (interviewer-interviewé), un nouveau tour de parole est créé. Le dernier type de découpage consiste en une segmentation qui se fait à l'intérieur d'un même tour de parole. Nous avons adopté le modèle de segmentation de Lefeuvre qui se base sur des critères syntaxiques. Selon elle, la phrase se compose d'une unité prédicative qui peut être soit autonome avec une modalité d'énonciation soir non autonome c'est à dire enchâssée (Lefeuvre, 2021). Nous nous sommes également basée sur des critères prosodiques en tenant compte de l'intonation et des pauses. Une fois les données transcrites, elles sont importées vers le logiciel TXM. Vu que ce type de mots n'appartient à aucune des parties du discours, une annotation manuelle s'impose. Voici les différentes étapes de l'annotation des MDs selon le modèle d'Abouda (2022) ainsi que le schéma d'annotation (Abouda, 2022) :

- **Étape 1** : Identifier les MDs qui nous intéressent
- **Étape 2** : Différencier l'emploi discursif de l'emploi non discursif
- **Étape 3** : Définir des catégories d'analyse : position, portée, valeur sémantico-pragmatique, substitution.
- **Etape 4** : Ajouter des valeurs pour chaque catégorie en fonction du marqueur analysé : Exemple : position : initiale, intermédiaire, finale / portée : contexte gauche, droit

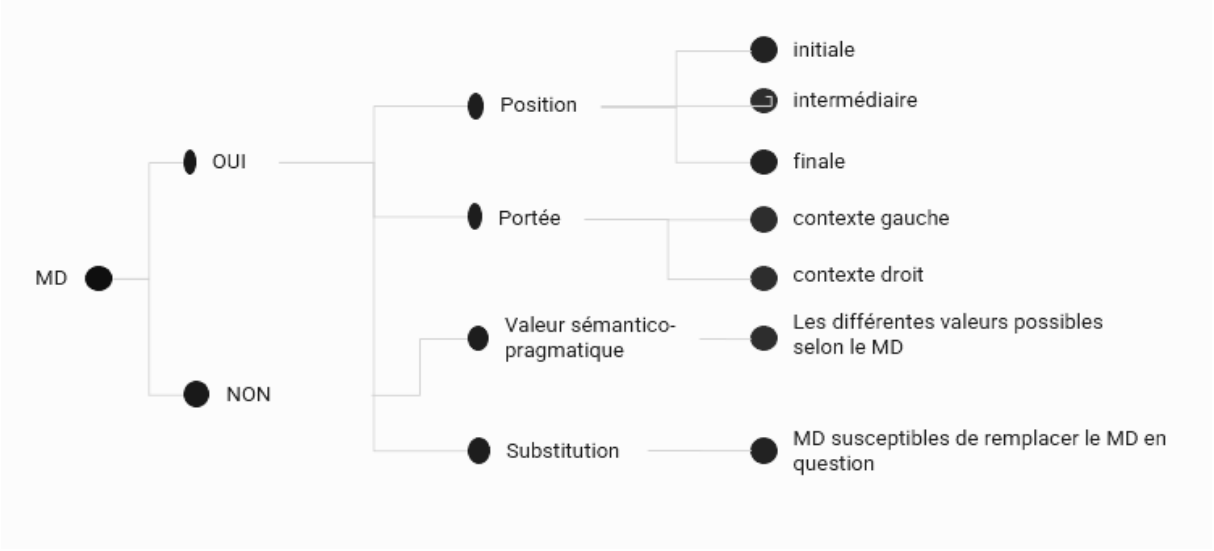


figure . 1 Figure 1 Schéma d'annotations des MDs

4 Conclusion

En guise de conclusion, nous avons présenté dans ce papier nos choix méthodologiques concernant la constitution, la transcription et l'annotation de notre corpus qui servira de

corpus pour notre thèse portant sur l'étude des marqueurs discursifs du français tels qu'employés par des locuteurs tunisiens. Il est important de mentionner que ce corpus est en construction et qu'il n'est pas encore finalisé. Notre objectif actuel est d'homogénéiser le corpus en tenant compte des différents paramètres extralinguistiques.

Références bibliographiques

Abouda, L. (2022). L'émergence du marqueur méta-discursif du coup : De la conséquence à l'actualisation énonciative. *Langages*, *226*(2), 99-116. https://doi.org/10.3917/ lang.226.0099

Abouda, L., & Baude, O. (2006). Constituer et exploiter un grand corpus oral : choix et enjeux méthodologiques. Le cas d'ESLO, CORAL-Université d'Orléans.

Baude, O., Blanche-Benveniste, C., Calas, M.-F., Cappeau, P., Cordereix, P., Goury, L., & Jacobson, M. (s. d). *Corpus oraux, guide des bonnes pratiques 2006*.

S. Branca‑Rosoff, S. Fleury, F. Lefeuvre, M. Pires. (2012) *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000*. Consulté 31 mai 2023, à l'adresse http://cfpp2000.univ-paris3.fr/CFPP2000

Dostie, G. (2004). Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique.Bruxelles : De Boeck et Duculot.

Dostie, G., & Pusch, C. D. (2007). Présentation. Les marqueurs discursifs. Sens et variation: *Langue française*, *n° 154*(2), 3-12. https://doi.org/10.3917/lf.154.0003

Guide Trancripteur (2013.). Consulté 31 mai 2023, à l'adresse http://eslo.huma-num.fr/images/eslo/pdf/GUIDE_TRANSCRIPTEUR_V4_mai2013

Lefeuvre, F. (2021). Analyse outillée du marqueur discursif bien sûr. L'Information grammaticale, n°170, p. 32-42.

Vincent, D. (1993). *Les ponctuants de la langue et autres mots du discours*. Québec, Nuit Blanche.

Extraction de contextes riches en connaissances à partir d'un corpus comparable de textes médicaux (français-arabe)

Rim Abouwarda
Univ. Grenoble Alpes, LIDILEM, F-38000 Grenoble, France
rim.abouwarda@univ-grenoble-alpes.fr

Mots clés : terminologie textuelle – terminologie computationnelle - corpus comparable – contextes riches en connaissances (CRC) – marqueurs linguistiques – langue de spécialité

Introduction

Les langues de spécialité suscitent actuellement un grand intérêt du fait que les sciences et les techniques connaissent une évolution continue, ce qui est reflété par les besoins linguistiques accrus. De nos jours, les recherches en langue de spécialité se basent principalement sur l'observation et la description des phénomènes linguistiques en corpus, et ce, grâce à l'abondance des ressources textuelles numériques et les outils servant à les traiter. (Frérot & Pecman, 2021).

La présente analyse, inscrite dans le cadre de notre projet de thèse, est basée sur l'étude d'un corpus comparable de textes médicaux (français-arabe) dans le but de répondre aux besoins d'ordre traductionnel dépendant du contexte. En effet, les sources d'informations auxquelles ont recours les traducteurs sont souvent les dictionnaires ou les glossaires offrant un nombre limité de contextes. Il s'ensuit donc que la consultation des bases de données terminologiques présentant des informations lexicales et conceptuelles est fortement recommandée.

Objectifs de l'étude

L'étude s'adresse aux traducteurs en tant qu'« utilisateurs prioritaires » de la terminologie (Cabré, 1998). Notre recherche s'est donc fixée comme objectif de fournir aux traducteurs une contextualisation des termes dans les deux langues source et cible leur permettant de mieux saisir le sens des termes et d'en faire un usage à bon escient.

Cadre théorique

L'approche adoptée, dans le cadre de notre étude, est la terminologie textuelle (Bourigault et Slodzian, 1999) associée à la terminologie computationnelle et compatible avec l'optique lexico-sémantique (L'Homme, 2020). L'apport de l'approche textuelle réside dans le fait d'avoir élaboré une méthodologie outillée intégrant la linguistique de corpus et le TAL. C'est une approche, soulignons-le, qui accorde une importance au fonctionnement réel des termes dans leurs contextes (Condamines, 2018). D'où l'intérêt de se pencher sur l'extraction de contextes riches en connaissances (CRC) que Meyer (2001) définit en tant que fragments de textes contenant des éléments utiles pour l'analyse conceptuelle.

Dans la littérature, plusieurs études se sont intéressées à l'identification et à l'extraction des CRC en corpus comparable. Notons, à ce propos, la thèse de Hmida (2017) consacrée à l'étude des patrons de connaissances pour l'identification des définitions dans des corpus comparables portant sur l'oncologie et la vulcanologie. Cette étude a mis l'accent sur l'instabilité des marqueurs de relations due à la polysémie. Hypothèse corroborée par d'autres travaux comme Condamines (2002) et Marshman (2014). De plus, les marqueurs conceptuels ont été également étudiés dans une perspective diachronique par Picton (2009) pour définir des marqueurs d'évolution servant à identifier des CRC dans le domaine spatial.

Bien que la littérature sur le sujet soit assez riche, la dimension contrastive axée sur la combinaison linguistique français-arabe n'a pas été explorée. Dans ce sillage, il nous semble donc intéressant de se pencher sur les marqueurs lexico-syntaxiques en langue arabe permettant d'identifier des CRC. Pour ce faire, et en se basant sur les études menées sur le sujet, nous avons pu formuler l'hypothèse qu'une variation au niveau des patterns linguistiques entre le français et l'arabe pourrait être observée.

Problématique

Notre recherche tentera donc de répondre à la problématique suivante : Comment, à partir de l'interrogation d'un corpus de textes comparable dans le domaine médical, extraire des contextes définitoires et identifier les marqueurs lexico-syntaxiques dans une perspective contrastive dans les deux langues français et arabe ?

Cadre méthodologique et corpus

Notre corpus comparable est constitué de textes médicaux (français-arabe) portant sur le domaine de la psychiatrie et la psychologie clinique.

Critères	Corpus français	Corpus arabe
Période	2000-2021	1980-2021
Type de documents	Articles scientifiques Manuels de psychiatrie	Ouvrages scientifiques Articles scientifiques
Degré de spécialisation	Rédigés par des experts Destinés à des spécialistes Destinés au grand public	Rédigés par des experts Destinés à des spécialistes Destinés au grand public
Taille	18 revues scientifiques (n° de 2000 à 2021) + 7 manuels de psychiatrie	20 ouvrages scientifiques + 5 revues scientifiques (n° de 2014 à 2021)
Format	XML et TXT	TXT

table 1. : table 1 : Constitution du corpus comparable

Pour le traitement outillé du corpus, nous utilisons le Lexicoscope (Kraif, 2019).

Quant à l'extraction des contextes définitoires en français, nous nous sommes basé sur la liste MAR-EL pour les marqueurs conceptuels (Lefeuvre, 2017). Pour ce qui concerne les contextes définitoires en arabe et vu l'absence d'une liste existante de marqueurs linguistiques, nous avons extrait et analysé les contextes où apparaissent les candidats-termes. Ces derniers ont été extraits grâce aux patterns productifs prédéfinis dans le cadre de notre thèse sur la base d'une liste de fréquence de référence, tant au niveau des lemmes que des suites de lemmes.

Il nous revient de noter que Lehmann et Martin-Berthet (2008) ont classé les types de contextes en trois catégories : définitoires, encyclopédiques et linguistiques. Nous nous focalisons, aux fins de notre analyse, sur les contextes définitoires renfermant des éléments permettant de se renseigner sur la signification du terme.

Analyses et résultats

Tout d'abord, les observations portées sur le corpus bilingue nous ont permis de constater des convergences dans l'emploi de certains marqueurs linguistiques comme :

Français	**Terme +**	**V. être +**	**Déterminant + Nom +**	**Caractérisé(e) +**	**par**
Arabe	**المصطلح+**	**هو+**	**اسم نكرة+**	**يتميز+**	**ب**

Exemple extrait du corpus français :

La schizophrénie est une affection caractérisée par un ensemble de signes particuliers, incluant des idées délirantes, des hallucinations, un discours désorganisé, (…)

Exemple extrait du corpus arabe :

اضطراب وجداني ثنائي قطبي هو اضطراب يتميز بنوبات متكررة يضطرب فيها مزاج الشخص ومستوى نشاطه بشكل عميق.

بشكل عميق	نشاطه	مستوى	و	الشخص	مزاج	يضطرب فيها	متكررة	نوبات	ب	يتميز	اضطراب	هو	ثنائي قطبي	وجداني	اضطراب
profondément	d'activité	Son niveau	et	de l'individu	l'humeur	qui troublent	récurrentes	des crises	par	caractérisé	un trouble	est	bipolaire	affectif	Le trouble

En outre, nous remarquons, dans le corpus arabe, une prédominance des indices typo-dispositionnel comme « : », tel que dans l'exemple suivant :

- اضطراب التوافق: هي حالات من الضيق الذاتي والضيق الانفعالي (...)

الانفعالي	الضيق	و	الذاتي	الضيق	من	حالات	هي	التوافق :	اضطراب
émotionnel	du malaise	et	personnel	malaise	de	des états	est	d'adaptation :	Le trouble

Un autre indice de type lexical a été attesté dans le corpus arabe, c'est l'emploi du mot « التعريف » (la définition) précédant le contexte définitoire :

- الهذيان الارتعاشي:

التعريف: هو مرض عقلي ذهاني ونوع خاص من الهذيان الحاد (...)

الحاد	الهذيان	من	خاص	نوع	و	ذهاني	عقلي	مرض	هو	التعريف:	الارتعاشي:	الهذيان
grave	délire	du	particulier	un type	et	délirante	mentale	une maladie	est	La définition :	tremens	Le delirium

Quant au corpus français, nous assistons à un emploi plus fréquent des indices lexicaux à savoir : « comme, c'est-à-dire » :

- L'autisme est défini comme une maladie complexe du développement du SNC et est associé à une étiologie multifactorielle.

- L'anorexie mentale chez les adolescentes se présente comme un syndrome qui associe simultanément ou progressivement : une perte d'appétit ; un amaigrissement ; l'arrêt des règles (aménorrhée) ; (…)

- La survenue d'attaques de panique régulières peut aussi conduire à l'installation d'une agoraphobie, c'est-à-dire la crainte de se trouver dans un lieu ou une situation d'où il sera impossible de s'échapper en cas de survenue d'une attaque de panique.

Notre communication sera donc structurée autour de deux axes. Dans un premier temps, nous aborderons le cadre théorique et la problématique de notre recherche. Ensuite, nous passerons en revue la méthodologie sur laquelle nous nous sommes basé dans le cadre de notre analyse. Dans un second temps, nous présenterons les résultats de notre étude étayés par des exemples extraits du corpus.

Références bibliographiques

Bourigault, D. & Slodzian, M. (1999). Pour une terminologie textuelle. *Terminologies nouvelles.* N°19, pp. 29-32.

Cabré, M-T. (1998). *La terminologie. Théorie, méthode et applications* (traduit par Cormier, M. et Humbley, J.). Ottawa : Les Presses de l'Université d'Ottawa. 322 pages.

Condamines, A. (2002). Corpus analysis and conceptual relations patterns. In *Terminology*. John Benjamins Publishing Company, pp. 141-162.

Condamines, A. (2018). Nouvelles perspectives pour la terminologie textuelle. In Altmanova, J., Centrella, M. & Russo, K.E (Eds). *Terminology and Discourse*. Peter Lang.

Frérot, C. & Pecman, M. (dir.) (2021). *Des corpus numériques à l'analyse linguistique en langues de spécialité*. Grenoble : UGA Éditions. 373 pages.

Hmida, F. (2017). *Identification et exploitation de contextes riches en connaissances pour l'aide à la traduction terminologique.* (Thèse de doctorat, Université de Nantes).

Kraif, O. (2019). Explorer la combinatoire lexico-syntaxique des mots et expressions avec le LEXICOSCOPE. In *Langue française*. N°203, pp. 67-82.

Lefeuvre, L. (2017). *Analyse des marqueurs de relations conceptuelles en corpus spécialisé : recensement, évaluation et caractérisation en fonction du domaine et du genre textuel*. (Thèse de doctorat, Université Toulouse Jean Jaurès).

Lehmann, A. & Martin-Berthet, F. (2008). *Introduction à la lexicologie : sémantique et morphologie* (3e édition revue et actualisée). Paris : Armand Colin. 261 pages.

L'homme, M-C. (2020). *La terminologie. Principes et techniques* (2e édition revue et mise à jour). Montréal : Presses de l'Université de Montréal.

Marshman, E. (2014). Enriching terminology resources with knowledge-rich contexts : A case study. In *Terminology*. Vol. 20. N°2, pp. 225-249.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In D. Bourigault, M-C l'Homme & C. Jacquemin (Eds). *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 279–302.

Picton, A. (2009). *Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial.* (Thèse de doctorat, Université Toulouse Le Mirail).

Enseignement sur corpus : conscience pragmatique et compétence interactionnelle

Carmen Alberdi [1] et Carole Etienne [2]
[1] Université de Grenade
[2] Laboratoire ICAR, CNRS / ENS Lyon
kalberdi@ugr.es, carole.etienne@ens-lyon.fr

Introduction

Acheter son pain à la boulangerie, accueillir des amis chez soi, inviter quelqu'un pour une soirée ou poser une question à un collègue : autant d'actes simples et quotidiens qui témoignent néanmoins de l'insuffisance d'une approche méthodologique qui ne tiendrait pas compte du besoin, pour l'apprenant, de développer, au fil de son apprentissage, une conscience pragmatique (Meier 2003 ; Pizziconi & Locher 2015), lui permettant d'acquérir une vraie compétence interactionnelle (André 2021).

En effet, les composantes pragmalinguistiques et sociopragmatiques de ce que le Cadre Européen Commun de Référence pour les Langues (CECRL, Conseil de l'Europe 2001) regroupe sous l'alinéa des sous-compétences sociolinguistiques font rarement l'objet d'un enseignement explicite. D'ailleurs, l'enseignement de l'oral prenant le plus souvent appui sur la langue écrite (Germain & Netten 2010), les valeurs pragmatiques que les expressions acquièrent en situation d'interaction demeurent, en général, largement confiées à l'intuition de l'apprenant. Plus les langues -maternelle et étrangère- sont proches (comme l'espagnol et le français, par exemple), plus l'apprenant est pris dans une sorte d'« illusion de transparence » (Kulikowski 2012) qui le pousse à un transfert de structures calquées de sa langue maternelle, au risque de tomber dans des erreurs pragmatiques (Thomas 1983 ; Kasper & Rose 2002). Or, en fonction du niveau de l'apprenant, celles-ci sont moins perçues comme un manque de compétence linguistique que comme un manque de politesse ou d'intérêt (Escandell 1996 ; Piatti 2003). L'utilisation d'outils et de ressources didactiques élaborées sur des corpus d'interactions s'avère, dans ce sens, un auxiliaire incontournable.

Ressources et méthodologie

Ressources

La présentation de données réellement attestées dans des situations sociales ordinaires permet d'illustrer en contexte la manière dont les activités se déroulent au quotidien (Traverso 2016) et comment elles s'enchaînent au cours de l'interaction en simulant, dès lors que l'exposition est renouvelée régulièrement, une forme d'immersion dans ces données (Alberdi et *al.* 2018). C'est dans ce but que l'équipe Interactions, Cognitions du laboratoire ICAR a conçu et développé la plateforme CLAPI-FLE (http://clapi.icar.cnrs.fr/FLE/), en collaboration avec un réseau de didacticiens et d'enseignants. Si cette plateforme propose une quarantaine d'extraits

didactisés pour l'enseignement du français langue étrangère, elle met également à disposition des fiches explicatives pour détailler certaines particularités de l'oral (temps verbaux, atténuation, discours indirect, questions, expressions "quand même" ou "trop") et, en parallèle, des fiches "En pratique" directement utilisables en salle de classe qui portent sur l'accueil, la cuisine, l'invitation (Alberdi et Etienne 2021) ou encore sur les différentes formes que peut prendre une question suivant les contextes et les besoins.

Orientée vers les enseignants (Ravazzolo et Etienne 2019), cette plateforme s'avère souvent complexe pour les apprenants qui ont des difficultés à appréhender ses contenus, en particulier le métalangage qui les décrit. Ainsi, une seconde plateforme dédiée aux apprenants CORAIL (http://clapi.icar.cnrs.fr/Corail) a été définie, dans ce même réseau, d'après les résultats d'une enquête qualitative menée auprès d'apprenants de différentes nationalités et de différents niveaux de langue. Elle est basée sur la compréhension, après plusieurs écoutes, d'extraits authentiques revue et discutée lors d'entretiens individuels afin de cibler les principales difficultés de compréhension des apprenants (Cortier et al. 2022).

À partir des problèmes recensées, l'application s'organise en trois volets : les situations, les fonctions et les expressions, que l'apprenant aborde dans l'ordre qui lui convient suivant ses besoins.

CORAIL propose une quinzaine d'extraits courts et commentés, accompagnés d'exercices qui pointent quelques spécificités de l'oral, souvent source de confusion.

figure . 1 CORAIL : Variété des situations privées ou professionnelles du quotidien

En parallèle, la rubrique "Comment dire ? Comment faire ?" décrit dans un langage simple, et illustre dans différents contextes, les procédés mobilisés par les locuteurs pour réaliser des fonctions courantes telles que remercier, raconter, poser une question, exprimer une opinion, proposer, inviter, mais également des actes plus délicats comme refuser ou plaisanter. Même si

CORAIL s'adresse à des apprenants qui peuvent être débutants, il ne s'agit pas de montrer une façon de faire, bien rodée et maitrisée, mais au contraire de donner accès à une variété de procédés afin de préparer les apprenants à les reconnaitre, à les assimiler pour développer de réelles compétences d'écoute et de repérage, dans le but d'améliorer ainsi leur compréhension de la langue orale.

Dans la même approche, certaines expressions fréquentes à l'oral et identifiées par les apprenants comme problématiques sont présentées, en abordant en termes simples leur polysémie dans des exemples contextualisés. En évitant l'écueil d'une traduction intrinsèque qui ne rendrait pas compte de la diversité de leurs emplois, on préfère décrire leurs fonctions, leurs variantes, leur caractère positif ou négatif, la prosodie ou les gestes qui les accompagnent. Le maillage entre lexique, fonction, cotexte, prosodie et contexte sensibilise les apprenants à élargir leur approche de la langue pour intégrer plusieurs de ses composantes, au lieu de se limiter à une connaissance parcellaire de quelques-uns de ces termes, réutilisables à l'identique et sans discernement.

Méthodologie

Nous illustrerons l'emploi et l'adaptabilité de ces ressources à travers une étude de cas menée au sein d'un groupe d'apprenants de français 2e langue étrangère de la Faculté de Traduction de l'Université de Grenade (3e année de licence, dernier semestre de langue française).

Bien qu'ayant suivi 360h de cours de français à l'université -sans compter les parcours individuels au long de l'enseignement secondaire-, et malgré leur niveau de compétence général (B2+), les apprenants témoignent de difficultés d'expression -parfois aussi de compréhension- lorsqu'ils sont confrontés à des interactions simples comme celles envisagées ci-dessus. L'observation quotidienne permet de constater la persistance d'erreurs pragmatiques découlant du calque de structures de la langue maternelle -impératif pour une requête, emploi systématique de « salut », par exemple-, la tendance à assimiler traits syntaxiques d'oralité et registre familier -notamment en ce qui concerne l'interrogation ou la dislocation- et des problèmes d'identification de la valeur pragmatique de divers marqueurs de structuration typiques de l'oral -voilà, du coup, quand même…

Afin d'y remédier et de proposer aussi à l'apprenant des outils qu'il pourra continuer à utiliser de façon autonome, des ressources complémentaires ont été élaborées à partir de celles proposées dans CLAPI-FLE et CORAIL et utilisées en accompagnement des matériaux déjà prévus dans le cursus. Les séquences sont explicitement orientées vers un triple objectif :

1) l'acquisition d'une conscience pragmatique et interculturelle, à travers l'observation et la réflexion ;
2) le développement de la compétence interactionnelle en compréhension comme en expression ;
3) l'encouragement d'une conscience métacognitive à travers l'exploration de ressources d'auto-apprentissage et l'élaboration d'outils de synthèse personnalisés (cartes conceptuelles et journal de bord).

Résultats

Les résultats de l'expérience menée ont pu être interprétés à travers la comparaison entre les réponses fournies à une enquête préalable en début de semestre et celles données à un formulaire élargi portant sur les mêmes contenus après 14 séances de cours sur corpus. Si la

première enquête demandait juste une mobilisation de connaissances acquises dans la réalisation d'actions simples ou à l'égard de marqueurs habituels de l'oral tels que « du coup » ou « voilà », la seconde devait s'accompagner d'une explication détaillée et argumentée des conclusions de l'étudiant une fois les activités effectuées, et individuellement enregistrées dans le journal de bord.

Le caractère multidimensionnel de la langue orale rend compte de la richesse de ses pratiques mais également de sa variété, dont un apprenant ne peut se dispenser s'il veut pouvoir la comprendre en contexte. Pour ce faire, il lui faudra développer des stratégies d'écoute et de repérage des indices verbaux et multimodaux. Si les chercheurs en interaction disposent de corpus écologiques et des résultats de leurs travaux, ils ne constituent pas pour autant des ressources directement utilisables pour l'enseignement-apprentissage d'une langue. Un travail de transposition s'impose pour tirer le meilleur parti des connaissances des uns et des besoins des autres, afin de concevoir ensemble des ressources pertinentes qui évolueront au fil des connaissances comme des enseignements.

Dans un travail continu d'ajustement des ressources aux besoins, les enseignants et les didacticiens partagent avec nous leurs expériences auprès de publics variés et pour différents objectifs d'enseignement, afin de modifier ou de compléter les ressources en ligne selon leur réception par les apprenants et leur adéquation aux formations dispensées. C'est par ces approches successives que les plateformes se consolident et tentent d'apporter à la didactique des langues les résultats des nouveaux travaux de recherche en interaction, de nouvelles données, ainsi que de nouveaux formats adaptés à la salle de classe et à ses contraintes.

La création de telles ressources ne fonctionne qu'avec l'engagement des didacticiens et des enseignants, impliqués depuis le choix des extraits des corpus de recherche à didactiser, ou des fonctions langagières à expliciter, jusqu'à l'interprétation des résultats obtenus auprès des apprenants. Les améliorations à apporter, comme les nouvelles ressources à créer, découlent directement des échanges entre les chercheurs et les praticiens.

Références bibliographiques

Alberdi, C., Etienne, C. (2021). Apprendre à interagir en classe de FLE : la situation d'invitation. *Bulletin Suisse de Linguistique Appliquée*, n. spécial, 2, 107-128.

Alberdi, C., Etienne, C., Jouin-chardon, E. (2020). Comprendre les spécificités du français oral par l'immersion virtuelle : un défi possible pour les apprenants. *Études linguistiques - didactique des langues,* 19-38.

André, V. (2021). Des corpus d'interactions dans la formation linguistique des migrants. *Savoirs*, 56, 77-96.

Conseil de l'Europe (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*, Paris, Didier.

Cortier, C., Etienne C., Ould Benali, N. (2022). De la didactique de l'oral à l'interaction : rétrospective, méthodes et mise en œuvre dans le contexte algérien. *Mélanges CRAPEL*, n.1, 43, 101-129.

Escandell, M.ª V. (1996). Los fenómenos de interferencia pragmática. *Didáctica del español como lengua extranjera*, 95-110. Madrid : UNED.

Germain, C. & Netten, J. (2010). La didactique des langues : les relations entre les plans psychologique, linguistique et pédagogique. *F. Neveu, V. Muni Toke, J. Durand, T. Klingler, L. Mondada & S. Prévost (éds.), Congrès Mondial de Linguistique Française - CMLF 2010*, 519-537. Institut de Linguistique Française.

Kasper, G. & Rose, K. (2002). *Pragmatic development in a second language*. Oxford : Blackwell.

Kulikowski, M.Z. (2012). Los estudios sobre cortesía verbal en español en el departamento de letras modernas de la Universidad de São Paulo-Brasil. *J. Escamilla & G. H. Vega (éds.) : Miradas multidisciplinares a los fenómenos de cortesía y descortesía en el mundo hispano*, 325-343. Barranquilla-Stockholm: Universidad del Atlántico-Université de Stockholm.

Meier, A.J. (2003). Posting the banns: a marriage of pragmatics and culture in foreign and second language pedagogy and beyond. *A. Martínez, E. Usó & A. Fernández (éds.): Pragmatic competence and foreign language teaching* ,185-210. Castellón : Université Jaime I.

Piatti, G. (2003). La cortesía : un contenido funcional para los programas de español como lengua extranjera. *D. Bravo (éd.) : La perspectiva no etnocentrista de la cortesía : identidad sociocultural de las comunidades hispanohablantes,* 355-368. Stockholm: Université de Stockholm -EDICE.

Pizziconi, B. & Locher, M.A. (éds.) (2015). *Teaching and learning (im)politeness*. Berlin : Mouton de Gruyter.

Ravazzolo, E., Etienne, C. (2019). Nouvelles ressources pour le FLE à partir des études en interaction. *LINX. Apprendre à interagir en langue étrangère, réflexion linguistiques et didactiques*. [https://journals.openedition.org/linx/3454]

Thomas, J. (1983). Cross-Cultural Pragmatic Failure, *Applied Linguistics*, 4, 91-112.

Traverso, V. (2016). *Décrire le français parlé en interaction*. Paris : Ophrys

Effect of Using Corpus-based Activities on Learning Certain Phrasal Prepositions among EFL Learners

Afnan Almegren [1]
[1] Department of Applied Linguistics, College of Languages, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

This paper explored the effect of using corpus-based activities on learning certain phrasal-prepositional verbs. The participants of the study were Saudi EFL learners. This study intended to investigate whether six hours of corpus-based activities instruction was efficient in teaching the forms of 40 phrasal-prepositional verbs. It also aimed to explore whether the assigned duration assisted the participants understanding of the metaphorical meaning of the specified prepositional verbs. And whether it assisted the correct form construction of the verbs while paraphrasing the sentences. The corpus-based activities that were used over six hours were indirect.

The activities used were obtained from the Corpus of Contemporary American English. They included forty phrasal-prepositional verbs. The outcome of this research revealed that six hours of instruction using corpus-based activities had a great impact on the form and use of phrasal-prepositional verbs among the specified EFL learners. The findings also present significant results in relation to learners' understanding of the phrasal- prepositional verbs metaphorical meaning.

Keywords: Corpus-linguistics, Corpus-Based Activities, Phrasal-prepositional Verbs, EFL Learners, Teaching EFL

This paper was published in *The International Journal of Communication and Linguistic studies* (SCOPUS ranking)

https://doi.org/10.18848/2327-7882/CGP/v20i01/1-11

La négociation dans l'enseignement de l'espagnol des affaires : Constitution d'un corpus à visée didactique

Sara Alvarez Martinez
ILCEA, Université Grenoble Alpes
Sara.alvarez@univ-grenoble-alpes.fr

Introduction

Dans le cadre de l'enseignement de l'espagnol des affaires, nous nous intéressons aux négociations en tant que genre de spécialité oral. Nous entendons par négociation « a process of communicatif interaction through which two or more parties aim to solve their conflicting interests in a way that all parties regard as preferable to any alternative (Bülow, 2009 : 142). La négociation constitue l'un des discours les plus utilisés dans les échanges professionnels de la vie réelle et, par conséquent, une technique de communication orale très répandue dans le domaine de l'espagnol pour objectifs spécifiques.

De nombreux travaux se penchent sur les *business negotiations* (négociations commerciales) montrant une grande variété d'approches, parmi lesquelles on peut citer les recherches prescriptives (Fisher, Ury & Patton, 1991), les études théoriques (Yong and Saeidi, 2012), les travaux ethnographiques (Friedman, 2004), les approches basées sur le discours (Tripp & Sondak, 1992) ou encore les études portant sur la théorie des genres. La négociation a ainsi été une question centrale dans les recherches en linguistique appliquée. Pourtant, les aspects linguistiques de ce type de discours ont été négligés dans la littérature. Gardani (2017) constate « a remarkable dearth of publications that focus on training students and practitioners to negotiate in a second or third language other than English ».

En effet, malgré l'intérêt que les négociations ont suscité depuis les années 90, il n'existe pas à l'heure actuelle de matériel didactique (manuel) sur les négociations adressé à des étudiants d'espagnol langue étrangère. Les manuels d'espagnol niveau B2 où ce type d'interaction fait partie de leurs tables de matières (par exemple : *Negocio a la vista* (2004), *Negocios. Manual de español profesional* (2005), *En equipo.es* 3 (2007), *La comunicación oral en la empresa* (2008), *Entorno empresarial* (2008), *Expertos* (2009), *Asuntos de negocios* (2010)) le font de façon très synthétique. En effet, même si on y retrouve de précieuses pistes du point de vue didactique, on constate un manque de données représentatives pour former de futurs négociateurs en espagnol langue étrangère.

Dans ce contexte, nous nous proposons de fournir une caractérisation de la négociation à partir de la théorie des genres en constituant un corpus de FASP (Fiction à substrat professionnel) sur la négociation à visée didactique. Nous nous penchons sur l'analyse de l'organisation rhétorique des négociations en précisant les *macro-moves (macro-séquences)*, les *moves* (séquences) et les *steps (étapes)* pour ensuite aborder les retombées didactiques dans l'enseignement de la négociation à des étudiants du Master LEA Négociation Trilingue en Commerce International.

Corpus et méthodologie

Corpus

Face aux difficultés de collecter des négociations réelles dans le domaine professionnel, en particulier pour des raisons de privacité et de confidentialité dans le monde professionnel, nous avons exploré d'autres sources pouvant offrir des données représentatives. Nous nous sommes orientés vers la fiction à substrat professionnel (FASP) car celle-ci peut constituer une source documentaire permettant de se faire une idée des réalités qui constituent le cadre de la langue et du discours de spécialité (Petit, 1999). Nous avons ciblé la *fiction des affaires* (Hardy, 2010) et, plus précisément, des films où nous avons repéré des extraits de négociations commerciales.

Nous avons sélectionné dix films appartenant au genre cinématographique du drame (voir tableau 1). Les extraits analysés représentent 120 minutes. Après un laborieux travail de transcription et de modélisation à l'aide de l'outil ELAN, nous avons analysé ces extraits à partir d'une grille pour définir les macro-séquences, les séquences et les étapes, selon l'approche de Swales, Biber et Parodi.

Titre du film	Année d'apparition	Pays de production	Directeur	Genre
Erin Brokovich	2000	Etats-Unis	Steven Soderberg	Drame biographique/ Action / Aventures
El informador	2000	Etats-Unis	Ben Younger	Drame / Suspense / Crime
Jobs	2013	Etats-Unis	Joshua Michael Strn	Drame biographique
La clave del éxito	2011	Etats-Unis	Doug Liman	Drame biographique
El fundador	2016	Etats-Unis	John Lee Hancok	Drame biographique
En busca de la felicidad	2006	Etats-Unis	Gabriele Muccino	Drame biographique
El Padrino I	1972	Etats-Unis	Francis Ford Coppola	Drame criminel
El Padrino II	1974	Etats-Unis	Francis Ford Coppola	Drame criminel
El método	2005	Espagne / Argentine	Marcelo Piñeyro	Drame
FIST Símbolo de fuerza	1978	Etats-Unis	Norman Jewison	Drame politique

table 1. : FASP Négociations commerciales

Méthodologie

Comme indiqué plus haut, nous prenons comme point de repère la théorie des genres de Swales (2004) et Biber, Connor & Upton (2007), reprise et revue plus tard par d'autres chercheurs hispanistes, comme Parodi (2009), coordinateur du projet CORPUS PUCV-2006 pour analyser de différents genres professionnels et académiques en espagnol dans quatre domaines de spécialité (psychologie, Travail Social, Ingénierie et Chimie Industrielle). Afin de caractériser la structure rhétorique des FASP Négociations, nous nous concentrerons sur la

notion d'objectifs communicatifs global et spécifiques ainsi que celle de fonction communicative en appliquant l'approche de l'analyse du discours orientée à la théorie (*Top-down approach*). Les étapes méthodologiques que nous avons suivies sont les suivantes :

1 Sélection des films en prenant en compte les critères du genre FASP.

2 Élaboration d'une grille d'analyse pour étudier la structure rhétorique de chaque négociation commerciale en nous servant du logiciel ELAN.

3 Identification des étapes (*steps*), des séquences (*moves*) et des macro-séquences (*macro-moves*).

4 Analyse et interprétation des données.

5 Transposition didactique des négociations commerciales analysées dans le cadre du Master LEA Négociation trilingue en commerce international.

Résultats

L'analyse réalisée du corpus de FASP Négociations commerciales nous a permis d'identifier 3 macro-séquences qui seraient : l'échange d'information, la discussion et le dénouement. Ces trois *macro-moves* (MM) peuvent se chevaucher, notamment si on n'arrive pas à un accord, et, dans certains extraits, être réduites à deux macro-séquences. Cela rend difficile de pouvoir réaliser une transposition didactique du processus complet de la négociation. En revanche, l'étude des séquences est particulièrement intéressante et riche pour la didactisation de la deuxième MM (la discussion), notamment à partir des séquences les plus fréquentes dans ce corpus : faire des propositions, argumenter un point de vue, faire des contrapositions, accepter ou refuser une proposition. Quant aux étapes détectées (*steps*) les plus fréquentes pouvant se prêter à une exploitation didactique enrichissante pour des futurs négociateurs, on peut citer : donner et demander des renseignements, persuader l'interlocuteur, faire un recommandation, donner un ordre, menacer, alerter ou prévenir l'autre partie.

Projection de l'étude

Cette étude sera complétée avec la constitution d'autres corpus de négociations réelles, de simulations de négociations (corpus d'apprenants) et de négociations pédagogiques (dans de manuels). Une analyse comparative de différentes négociations sera réalisée pour préciser la structure rhétorique dans les différents corpus dans le but de repérer les similitudes et les différences concernant l'identification des *macro-moves*, des *moves* et des *steps* pour ensuite créer du matériel didactique offrant des données plus représentatives et réelles de celles qu'offrent les manuels sur l'espagnol des affaires disponibles dans le marché actuellement.

Références bibliographiques

Biber, Douglas, Connor, Ulla et Upton, Thomas. (2007) (eds.) . (2007). *Discourse on the Move: Using Corpus Analysis to describe Discourse Structure*, Amsterdam: John Benjamins.

Bülow, Anne M. (2009). Negociation studies. In Francesca Bargiela-Chiappini (ed.), *The handbook of business discours*, 142-154, Edinburg : Edinburg University Press.

Fisher, Roger, William Ury & Bruce Patton (1991). *Getting to yes : Negotiating an agreement without giving in 2nd edn*, New York : Penguin Books.

Friedman, Ray (2004). « Studying negotiations in context : An Ethnographic approach », *International Negotiation* 9(3), 375-384.

Gardani, Francesco (2017), Business negotiations. In Gerlinde Mautner & Franz Rainer (eds.), *Handbook of Business Communication. Linguistic approaches*, 91-109.

Hardy, Mireille (2010), « Business FASP : un genre impossible ? », *ILCEA* [En ligne], 12 | 2010, mis en ligne le 23 septembre 2010.

Johns, Tim (2002). *Data Driven Learnin : The perpetual Challenge*. In Teachin and learning by Doing Corpus Analysis, 105-117.

Parodi, Giovanni (2008). « Lingüística de corpus : una introducción al ámbito », *Revista de Lingüística Teórica y Aplicada*, Concepción (Chile), 46 (1).

Pardi, Giovanni (2009): « El género manual y su organización retórica en cuatro disciplinas científicas: entre la abstracción y la con- creción », In Giovanni Parodi (ed.), *Géneros académicos y géneros profesionales. Valparaíso*: eUVSA, 119-228.

Petit, Michel (1999), « La fiction à substrat professionnel : une autre voie d'accès à l'anglais de spécialité », *Asp* (En ligne), 23-26.

Tripp, Thomas & Harris, Sondak (1992), « An evaluation of dependent variables in experimental negotiation studies : Impasse rats and pareto efficiency ». *Decision Processes in Negotiation* 51(2), 273-295.

Tribble, C. & G. Jones (1997), *Concordances in the Classroom: A Resource Book for Teachers* (second edition), Houston: Athelstan

Yong, Zhang & Sayedeh P. Saeidi (2012). The study on international business negotiation strategy based on incomplete information. In Yanwen Wu (ed.), *Advanced Technology in Teaching – Proceedings of the 2009 3rd International Conference on Teaching and Computational Science (WTCS 2009)*, 141-146 ; Berlin : Springer.

« Je comprends pourquoi mes amies françaises disent que je parle comme un livre ». Des corpus d'interactions endolingues et exolingues pour améliorer des compétences à l'oral en Français Langue Étrangère.

Virginie André[1] et Florence Poncet[1]
[1] Laboratoire ATILF, Université de Lorraine
Virginie.Andre@univ-lorraine.fr, Florence.Poncet@univ-lorraine.fr

Introduction

L'étude que nous présentons s'inscrit dans le cadre de l'exploitation des corpus en didactique des langues. Plus précisément, nous proposons d'utiliser des corpus multimodaux à des fins d'enseignement et d'apprentissage du français parlé en interaction (Ravazzolo *et al.*, 2015 ; André, 2020, 2019). Dans la lignée de l'utilisation des documents authentiques (Duda *et al.*, 1972 ; Abé *et al.*, 1979 ; Holec, 1990), nous proposons de recourir à des corpus d'interactions réelles afin d'exposer les apprenants au français parlé en usage ainsi qu'à la variété et à la variation des usages en situation. Nous nous appuyons sur deux façons d'exploiter ces corpus : en analysant les ressources verbales et non verbales mobilisées par les locuteurs pour arriver à leurs fins, c'est-à-dire en tirant profit des résultats des recherches en analyse des interactions (Traverso, 2016) et en les interrogeant avec un concordancier, selon les principes du *data-driven learning* (Johns, 1991) ou, en français, de l'apprentissage sur corpus (ASC) (Boulton, Tyne 2014). Plusieurs études montrent l'efficacité de ces approches didactiques, notamment pour l'enseignement et l'apprentissage de l'anglais écrit (Boulton, Cobb, 2017 ; Boulton, Vyatkina, 2021) mais également pour le développement de compétences orales et interactionnelles en français (Surcouf, Ausoni, 2018 ; Alberdi *et al.*, 2018 ; André, 2020 ; Etienne *et al.*, 2022).

Dans le cadre de cette étude, nous proposons de rendre compte d'expérimentations, réalisées au Département de FLE (DéFLE) de l'Université de Lorraine, visant à améliorer les compétences langagières d'apprenants de niveau intermédiaire, grâce à l'exploitation de corpus. L'objectif, pour ces apprenants, étaient d'apprendre à produire un récit en interaction, en observant et en analysant des locuteurs en train de raconter une anecdote ou une expérience. Par ailleurs, dans le cadre de cette même étude, nous avons souhaité examiner les exploitations, différenciées ou non, d'un corpus d'apprenants, non natifs du français, et d'un corpus de natifs. Pour ce faire, nous avons exploité le corpus multimodal du dispositif FLEURON (Français Langue Étrangère Universitaire : Ressources et Outils Numériques, https://fleuron.atilf.fr/). Ce dispositif numérique d'apprentissage du FLE met à disposition des apprenants et des enseignants un corpus multimodal d'interactions, outillé d'un concordancier

(André, 2016). FLEURON est dédié aux étudiants étrangers qui veulent se préparer à interagir avant de faire un séjour universitaire en France ou qui sont déjà en mobilité dans une université française.

Corpus et méthodologie

Corpus

Le corpus FLEURON tente de présenter les situations de communication auxquelles un étudiant étranger est susceptible de participer. Il compte actuellement un peu plus de 14 heures d'enregistrements vidéos d'interactions de deux natures différentes : 1) des interactions non sollicitées, qui ont été recueillies *in situ*, de façon écologique, et qui auraient eu lieu même si nous n'avions pas été présents pour les filmer ; 2) des interactions sollicitées, pendant lesquelles, sous la forme d'une conversation, des locuteurs expliquent des fonctionnements, des aspects culturels ou des éléments de la vie universitaire. Le corpus est classé selon des catégories pour aider les apprenants à se repérer dans les données. La table 1 présente les ressources du corpus[1], par catégories et selon la nature endolingue ou exolingue de l'interaction, c'est-à-dire si elle compte uniquement des locuteurs natifs ou au moins un locuteur non natif[2].

figure . 2 **Catégories**		figure . 3 **Endolingue**		figure . 4 **Exolingue**		figure . 5 **Total**	
figure . 6	Démarches administratives	figure . 7	35	figure . 8	6	figure . 9	41
figure . 10	Démarches à la préfecture	figure . 11	4	figure . 12	3	figure . 13	7
figure . 14	Étudiants Erasmus	figure . 15	7	figure . 16	18	figure . 17	25
figure . 18	Aides sociales	figure . 19	5	figure . 20	0	figure . 21	5
figure . 22	Questions pédagogiques	figure . 23	12	figure . 24	4	figure . 25	16
figure . 26	Santé	figure . 27	38	figure . 28	0	figure . 29	38
figure . 30	Explications du système universitaire	figure . 31	2	figure . 32	0	figure . 33	2
figure . 34	Utiliser les transports	figure . 35	13	figure . 36	1	figure . 37	14
figure . 38	La vie sur le campus	figure . 39	57	figure . 40	3	figure . 41	60
figure . 42	Témoignages	figure . 43	7	figure . 44	45	figure . 45	52
figure . 46	Culture et lieux de culture	figure . 47	20	figure . 48	0	figure . 49	20
figure . 50	Étudiants en doctorat	figure . 51	9	figure . 52	5	figure . 53	14

[1] Le corpus évolue régulièrement, ainsi que les catégories, selon les besoins des apprenants et des enseignants.
[2] FLEURON étant un dispositif d'apprentissage du français, les locuteurs non natifs du corpus sont des « étrangers compétents » (André, Castillo, 2005).

figure . 54	La vie en dehors du campus	figure . 55	41	figure . 56	4	figure . 57	45
figure . 58	**Total**	figure . 59	**250**	figure . 60	**89**	figure . 61	**339**

table 2. : Présentation du corpus FLEURON (en nombre de ressources, au 16/02/2023)

Les interactions du corpus sont toutes sous-titrées et transcrites (figure 1). Ces deux éléments peuvent être affichés ou non, simplement en cliquant sur un bouton.

figure . 62 Ressource du corpus FLEURON

Méthodologie

Les expérimentations se sont déroulées en 6 étapes. 1. Les apprenants ont visionné une interaction (non authentique) d'un manuel de FLE (avec la transcription). 2. Ils ont visionné 4 interactions du corpus FLEURON (2 endolingues, 2 exolingues). 3. Ils ont dû relever les spécificités du français parlé en interaction dans ces vidéos. 4. Ils ont recherché certaines de ces spécificités dans le concordancier[3] pour analyser et saisir leurs apparitions (voir l'exemple de « quoi » ci-dessous, figure 2). 5. Ils ont amélioré la transcription de l'interaction du manuel (étape 1) avec ces éléments. 6. Ils ont répondu à un entretien ou à un questionnaire concernant les interactions, les activités proposées et leur rapport à l'oral.

[3] Le concordancier de FLEURON permet de faire des allers-retours entre les occurrences et les interactions dans lesquelles elles sont prononcées.

16	E2: pour le reposer ou	quoi	E1: ouais
17	E1: ok bah merci pour ces informations E2: bah il y a pas de	quoi	
18	A : c'est compliqué cette année mais c'est comme ça on n'y peut pas grand chose	quoi	E1 : ouais
19	L1 : faut prendre un vol direct	quoi	L2 : non non non L1 : on peut même pas
20	A : oui ça fait à peu près une heure de permanence on va dire par semaine	quoi	
21	A: ouais c'est alors c'est en	quoi	
22	A : d'accord mais vous faites la langue française dans vers quel objecif votre master vous prépare à faire	quoi	
23	E1 : ça dépend A : oui dites moi ça dépend de	quoi	
24	E: je sais pas si par exemple je vous dit la métonymie vous savez pas c'est	quoi	
25	E1: tu voilà E2: voilà E1: tu leur dis ils notent et puis voilà après c'est bon	quoi	
26	E2: fiches ça sert à	quoi	ça E1: alors les petites fiches en fait c'est
27	E2: ben si tu as besoin de	quoi	que ce soit tu me dis E1: ok eh ben ça marche pas de souci je retiens
28	E: d'accord alors ça c'est pratique A: et à un prix défiant toute concurrence	quoi	

figure . 63 Extrait du résultat de la recherche de « quoi » via le concordancier de FLEURON

Ces expérimentations ont été menées en salle de classe, en décembre 2022 et en février 2023, avec deux groupes, de 8 et 11 participants. Ces séances ont duré respectivement 4 et 3 heures.

Résultats

Les résultats de ces expérimentations sont de différentes natures et concernent à la fois les données et la méthodologie didactique mise en œuvre. En voici quelques-uns, non hiérarchisés :

- Les apprenants qui interagissent avec des natifs comprennent immédiatement l'intérêt d'exploiter le corpus FLEURON.
- L'observation et l'analyse des interactions ne font pas partie des habitudes des apprenants.
- L'activité a permis la compréhension de spécificités de l'oral, notamment des marqueurs (*quoi*, *en fait*, *tu vois*, etc.). Certains se sont donnés comme tâche de les produire dans leurs prochaines interactions.
- Le dialogue de manuel est inégalement apprécié par les apprenants, selon leur culture d'apprentissage et selon leur degré d'intégration dans un environnement francophone.
- Les corpus exolingues sont également inégalement appréciés par les apprenants.
- Certains apprenants découvrent la « vraie » langue, éloignée de celle qu'ils ont apprise, et une d'entre eux explique : « je comprends pourquoi mes amies françaises disent que je parle comme un livre ».

Dans notre intervention, nous proposons de présenter plus précisément la méthodologie associée à la didactique de corpus et détaillerons tous les résultats de cette étude qualitative. De plus, ceux-ci seront complétés par une étude quantitative, en cours, sur la perception des interactions exolingues pour développer des compétences interactionnelles.

Références bibliographiques

Abé, D., Carton, F., Cembalo, S. M., et Régent, O. (1979). Didactique et authentique : du document à la pédagogie. *Mélanges pédagogiques*, 1-14.

Alberdi, C., Etienne, C., Jouin-Chardon, E.,2018, Les apports des corpus d'interactions naturelles en situation de classe : enjeux et pratiques, *Action Didactique*, 1, 55-70.

André, V. (**2020**). Corpus d'interactions et apprentissage du français langue étrangère. In Benzitoun C., Rebuschi M. *Les corpus en sciences humaines et sociales*. Nancy, Presses Universitaires de Nancy, 101-121.

André, V. (**2019**). Des corpus oraux et multimodaux authentiques pour acquérir des compétences sociolangagières. *In* Gajo L., Luscher J.-M., Racine I., Zay F. (eds), *Variation, plurilinguisme et évaluation en français langue étrangère*. Bern, Peter Lang, 209-223.

André, V. (2016). FLEURON : Français Langue Étrangère Universitaire – Ressources et Outils Numériques. Origine, démarches et perspectives. *Mélanges Crapel*, 37, 69-92.

André, V. & Castillo, D. (2005). The 'Competent Foreigner': A new model for foreign language didactics? In Preisler, B., Fabricius, A., Haberland, H., Kjaerbeck, S. & Risager, K., *The Consequences of Mobility*. Roskilde University, Department of Language and Culture, 154-162.

Boulton, A., Cobb, T. (2017). Corpus use in language learning: a meta-analysis. *Language Teaching*, 67, 348-393.

Boulton, A., Tyne, H. (2014). Des Documents Authentiques aux Corpus. Démarches pour l'Apprentissage des Langues. Paris, Didier.

Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology*, 25/3, 66-89.

Duda R., Esch E. et Laurens J. P. (1972). Documents non didactiques et formation en langues. *Mélanges pédagogiques,* 1-48.

Etienne, C., André, V. et Divoux, A. (2022). Interagir en réunion de travail : de l'étude des pratiques aux ressources didactiques. *CMLF 2022. SHS Web of Conferences*, 138.

Holec, H. (1990). Des documents authentiques, pour quoi faire. *Mélanges pédagogiques*, 65-74.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. In Johns, T., King, P. (dir.). *Classroom Concordancing, English Language Research Jounral*, 4, 1-16.

Ravazzolo, E., Traverso, V., Jouin, E., Vigner, G. (2015). *Interactions, dialogues, conversations : l'oral en français langue étrangèr*e. Paris, Hachette.

Surcouf, C., Ausoni, A. (2018). Création d'un corpus de français parlé à des fins pédagogiques en FLE : la genèse du projet FLORALE. *Études en didactique des langues*, 31, p.71-91.

Apports possibles des corpus au matériel pédagogique en FLE : une étude sur les requêtes dans les corpus écrits de natifs et d'apprenants

Sülün Aykurt-Buchwalter [1], Tatiana Aleksandrova [2]
[1] Laboratoire LIDILEM, Univ. Grenoble Alpes
[2] Laboratoire LIDILEM, Univ. Grenoble Alpes
sulun.aykurt@univ-grenoble-alpes.fr, tatiana.aleksandrova@univ-grenoble-alpes.fr

Introduction

La requête constitue un acte de langage défini comme « une demande d'un faire » (Kerbrat-Orecchioni, 2001 : 84). Dans des contextes sociaux authentiques comme dans des situations didactiques, les scripteurs en langue étrangère (L2) sont amenés à formuler des requêtes écrites. Cet acte de langage est fréquemment étudié dans les recherches en didactique des L2, plus particulièrement à travers l'angle de la pragmatique (Bardovi-Harlig, 1999). Les recherches portant sur la formulation des requêtes en français L2 indiquent que les scripteurs natifs et les scripteurs apprenants ne formulent pas les requêtes de la même manière, les natifs privilégiant les structures phraséologiques et les apprenants ayant tendance à employer une grande diversité de formes linguistiques (Warga, 2005). Des travaux sur les locuteurs avancés du français L2 concluent que les requêtes des quasi-bilingues contiennent des insuffisances pragmalinguistiques (Lundell, 2014). Dans l'enseignement/apprentissage du français langue étrangère (FLE), la requête écrite est souvent introduite au niveau B2, dans le cadre de l'étude du genre de la lettre formelle.

Les recherches sur l'exploitation pédagogique des corpus montrent que ces derniers constituent une ressource essentielle pour enrichir le matériel pédagogique. Ces apports se concentrent généralement sur l'enseignement/apprentissage du lexique, notamment les structures phraséologiques et les collocations (Cavalla, 2019). Le matériel pédagogique élaboré à partir des corpus est également le plus souvent axé sur le lexique : dictionnaires d'apprenants, listes de mots. Cependant, les corpus d'apprenants sont rarement exploités dans le matériel pédagogique, ce qui prive les apprenants d'une précieuse source d'information (Gilquin, 2007).

Dans le cadre de cette étude, nous essayons de comprendre si les résultats d'analyses menées sur des corpus écrits pourraient permettre d'améliorer ou de compléter le contenu des méthodes de FLE en ce qui concerne l'enseignement des requêtes écrites.

Corpus et méthodologie

Afin de répondre à cette question, nous procédons dans un premier temps à l'analyse contrastive d'un corpus de scripteurs francophones natifs et d'apprenants de FLE. Dans un deuxième temps, nous observons les éléments liés aux requêtes dans trois méthodes de FLE et deux ouvrages portant sur la production écrite en FLE au niveau B2 : *Edito B2* (Didier) ; *Alter*

Ego+ B2 (Hachette) ; *Echo B2* (CLE International) ; *Expression écrite niveau 4* (CLE International) ; *Production écrite niveaux B1/B2* (Didier).

Corpus d'essais argumentés

Le corpus est constitué d'essais argumentés rédigés d'une part par des étudiants francophones natifs et d'autre part par des étudiants apprenants de FLE ayant différentes langues initiales (L1). Ainsi, nous disposons de 20 essais de francophones natifs, 17 productions d'apprenants dont la L1 est l'indonésien, 20 productions d'apprenants russophones, et 15 productions d'apprenants turcophones. Notre corpus comporte donc 72 textes rédigés en français L1 et L2 sur le même sujet :

> *Vous habitez dans une ville qui organise chaque année un grand concert gratuit pour marquer la fin de l'été. Pour des raisons financières, votre ville annonce qu'elle veut supprimer cet événement musical. Vous écrivez au maire de la ville pour le persuader, à l'aide d'arguments et d'exemples précis, des avantages culturels et touristiques que ce concert représente. Vous insistez également sur l'intérêt économique de cette manifestation pour les commerçants et les artistes de la région. (250 mots).*

Nous avons veillé à contrôler, dans la mesure du possible, les profils de nos participants. Il s'agit dans tous les cas d'étudiants en licence et en master dans des filières de sciences humaines et sociales en France, en Indonésie, en Russie et en Turquie. Les participants ont rempli un questionnaire socio-biographique qui nous informe, entre autres, sur leurs pratiques linguistiques. Nous constatons donc qu'ils ont tous appris l'anglais au cours de leur cursus aux niveaux B1-B2, tout en sachant qu'ils ne pratiquent pas cette langue au quotidien et ne vivent pas dans un contexte bilingue.

Méthodologie

Notre analyse de corpus porte sur les points suivants :

- Des requêtes explicites sont-elles présentes dans les textes ? Dans quelle partie du texte sont-elles situées ? Quelle est la longueur de la requête en nombre de mots ?
- Quelles formes linguistiques sont-elles le plus souvent utilisées dans les requêtes ? En particulier, les occurrences de l'impératif, du verbe modal « pouvoir », des verbes volitifs « vouloir », « souhaiter », « demander » et « prier de » et des structures phraséologiques sont comparées.

L'analyse des manuels porte sur deux aspects :

- La requête constitue-t-elle un objet d'enseignement ? Si oui, dans quels contextes ?
- Quelles formes linguistiques sont employées dans la présentation des requêtes ?

Résultats

Les analyses du corpus révèlent que la requête est explicitement formulée dans environ 75% des textes analysés et se trouve essentiellement dans la partie finale du texte. Globalement, les requêtes sont plus nombreuses et plus longues chez les scripteurs francophones natifs que chez les apprenants. Les analyses du corpus montrent que les moyens employés pour exprimer la requête en français L1 et L2 sont variés. Les francophones natifs privilégient la forme « je vous prie de » pour formuler leur requête. Cette forme représente 46% des moyens utilisés par ce groupe. En revanche, les apprenants n'emploient que rarement cette formulation et optent majoritairement pour le verbe « espérer » à la première personne du singulier : « j'espère que ».

Une autre caractéristique des scripteurs apprenants - en particulier russophones et turcophones - est la surutilisation de verbe « demander » (« je vous demande de/ nous vous demandons de »). Les apprenants indonésiens se distinguent de cette tendance et optent davantage pour les expressions « il serait souhaitable de… » et « si vous voulez bien ». Les apprenants turcophones surutilisent le verbe « pouvoir » dans la formulation des requêtes par rapport aux natifs. On observe donc que les scripteurs apprenants se distinguent des scripteurs natifs dans la formulation des requêtes, ce qui confirme les recherches existantes. Certaines caractéristiques sont propres à certains groupes d'apprenants ; d'autres sont communes aux trois groupes scripteurs en L2.

L'analyse des manuels, quant à elle, indique que la requête ne constitue un objet d'enseignement que dans une méthode de FLE. Elle est présente, en revanche, dans les deux ouvrages portant sur la production écrite. Les formes linguistiques, présentées dans le cadre de lettres formelles non authentiques ou de manière isolée, sont relativement variées. Elles permettent d'exprimer des requêtes plus ou moins directes, allant du verbe « solliciter » au verbe « pouvoir », mais aussi la forme interrogative, les verbes « souhaiter » et « prier de », le conditionnel, et les structures phraséologiques « je vous serais reconnaissant(e) de » ou « auriez-vous l'obligeance de ».

Ainsi, il existe des points communs entre les formes employées par les scripteurs natifs de notre corpus et les formes proposées dans le matériel pédagogique étudié. Les deux points où le matériel pédagogique semble insuffisant sont : (1) l'absence de la requête dans les méthodes et (2) la prise en compte des difficultés des apprenants et de leurs erreurs fréquentes. Il serait alors pertinent d'une part d'inclure la requête comme objectif d'enseignement/apprentissage transversal à travers différents genres de textes, et d'autre part, de contraster, dans le matériel pédagogique, les formes correctes avec les formes à éviter, tout en proposant des exercices de reformulation en tenant compte d'un changement de destinataire.

Références bibliographiques

Bardovi-Harlig, K. (1999). Exploring the Interlanguage of Interlanguage Pragmatics: A Research Agenda for Acquisitional Pragmatics. *Language Learning*, *49*(4), 677–713. https://doi.org/10.1111/0023-8333.00105

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, *6*(4), 319–335. https://doi.org/10.1016/j.jeap.2007.09.007

Kerbrat-Orecchioni, C. (2001). *Les actes de langage dans le discours*. Nathan.

Forsberg Lundell, F. (2014). Demander un service en français L2 quasi-natif : Aspects pragmalinguistiques et socio-pragmatiques. *Synergies Pays Scandinaves*, (9), 93–107

Warga, M. (2005). "Je serais très merciable": Formulaic vs. Creatively Produced Speech in Learners' Request-Closings. *Canadian Journal of Applied Linguistics*, *8*(1), 67–93.

Méthodes de FLE

Berthet. A., & al. (2015). *Alter Ego+ B2*. Hachette.

Dupleix, D., Maigre, B. (2007). *Production écrite niveaux B1-B2*. Didier.

Girardet, J., Gibbe, C. (2013) *Écho 2e édition B2*. CLE International.

Heu, E. (2016). *Edito B2*. Didier.

Poisson-Quinton, S., Mimran, R. (2017). *Expression écrite B2*. CLE International.

Analyse outillée de corpus d'interactions de classe : prendre en compte les individualités au sein des interactions

Sophie Babault[1]
[1]Université de Lille
s.babault@yahoo.fr

Inscrite dans la thématique "corpus et didactique" du colloque JLC 2023, ma communication abordera d'un point de vue principalement méthodologique la prise en compte des individualités dans l'analyse de corpus d'interactions de classe.

De nombreuses recherches ont mis en avant le rôle joué par la médiation langagière dans la construction des savoirs scolaires (Babault, 2015 ; Gajo, 2007 ; Jaubert, 2007 ; Schneeberger & Vérin, 2009 ; etc.). La notion de médiation langagière est ici comprise comme un ensemble de « transactions [sociales] utilisant le médium verbal » (Bronckart, 2004).

Au sein des recherches sur la médiation langagière intervenant dans l'enseignement-apprentissage des disciplines scolaires, les interactions verbales occupent une place centrale et ont fait l'objet d'analyses explorant la relation entre les interactions verbales et le savoir (Duff, 2002 ; Malkoun & Tiberghien, 2008; Jaubert & Rebière, 2012), la co-construction discursive (Fillietaz & Schubauer-Leoni, 2008 ; Nonnon, 2018), les pôles enseignant-élèves (Broussal & Boucheton, 2008), etc.

La majorité de ces études construisent un sujet élève s'inscrivant dans un collectif qui laisse à la marge les modalités d'articulation des interactions de classe avec les démarches individuelles de chacun. Ainsi, chaque élève pourra être considéré comme un membre de la communauté discursive, comme l'indiquent souvent les transcriptions d'enregistrements : élève 1, élève 2, etc. Cependant, en dépit de cette présentation individualisée des transcriptions, les analyses mettent généralement en avant l'apport de chacun à la construction collective, plutôt que des cheminements discursifs individuels. Ce constat m'a conduite à questionner de manière dialectique la relation entre construction discursive et notionnelle plurielle versus construction individuelle. Comment la co-construction discursive attestée lors d'interactions de classe, mais n'impliquant pas nécessairement tous les élèves de la même manière, s'articule-t-elle avec la nécessaire part d'individualité des processus de construction discursive et conceptuelle ?

Si la pertinence de cette problématique semble ne pas faire de doute, les modalités méthodologiques de son traitement constituent des sources considérables d'interrogation à chaque étape de la démarche scientifique :

- recueil de données lors des interactions de classe : nécessité d'un traçage précis et rapide des prises de parole de chaque élève, le seul enregistrement audio ou vidéo ne suffisant pas dans le cadre d'une classe d'environ 25 élèves;

- transcription des données et constitution du corpus: nécessité de suivre chaque élève au fil des différents enregistrements, l'unité de temps d'un enregistrement de classe étant généralement un cours de 45 à 90 minutes, tout en respectant l'anonymisation du corpus ; nécessité d'assurer ce suivi en respectant l'éventuelle diversité des corpus (suivi des élèves dans des disciplines différentes, enregistrements de classe couplés à d'autres types de données recueillies, etc.);
- analyse des corpus: nécessité de prévoir un mode d'identification et de suivi des élèves compatible avec une analyse outillée, incontournable pour un corpus volumineux; complexité de l'annotation en raison de la diversité des formes d'implication de chaque élève dans la construction discursive ;
- etc.

En s'appuyant sur les recherches menées dans le cadre du projet INTERDID (Interactions didactiques et construction des savoirs scolaires), ma communication aura pour objectif de détailler l'ensemble de ces contraintes méthodologiques et les réponses qui peuvent y être apportées avec un certain degré d'efficacité. Le corpus est constitué, d'une part, d'interactions de classe en 4e dans des collèges français (8e année de scolarisation) et, d'autre part, d'entretiens individuels menés avec ces élèves de 4e à l'issue d'une série de plusieurs séances. La transcription des interactions et entretiens est en cours au moyen du logiciel CLAN (Childes). Les analyses textuelles seront réalisées avec le logiciel Hyperbase (BCL).

Références bibliographiques

Babault, S. (2015). *L'éducation bilingue : enjeux de politique linguistique, appropriation par les acteurs sociaux, développement de compétences chez les apprenants*. Habilitation à diriger des recherches, document de synthèse. Université Lille 3.

Broussal, D., Bucheton D. (2012). Interagir en début de cours : enjeux didactiques et discursifs . *Éducation et didactique* [En ligne], vol 2 - n°3 | Décembre 2008, mis en ligne le 01 décembre 2010. URL : http://educationdidactique.revues.org/357

Bronckart, J.P. (2004). La médiation langagière. Son statut et ses niveaux de réalisation. In R. Delamotte (dir.), *Les médiations langagières. Vol. II, Des discours aux acteurs sociaux*. Rouen : PUR, 2004. p. 11-32

Duff, P. A. (2002). The discursive co-construction of knowledge, identity, and difference: Ethnography of communication in the high school mainstream. *Applied Linguistics, 23*(3), 289–322. https://doi.org/10.1093/applin/23.3.289

Fillietaz, L., Schubauer-Leoni, M.L. (2008). *Processus interactionnels et situations éducatives*. Bruxelles : De Boeck Supérieur.

Gajo, L. (2007). Enseignement d'une DNL en langue étrangère : de la clarification à la conceptualisation. *TREMA*, 28, 37-48.

Jaubert, M. (2007). *Langage et construction de connaissances à l'école : un exemple en sciences.* Bordeaux : Presses universitaires de Bordeaux.

Jaubert, M., Rebière, M., Pujo, J. (2010). Communautés discursives disciplinaires scolaires et formats d'interaction. Colloque international "Spécificités et diversité des interactions didactiques : disciplines, finalités, contextes", Université de Lyon - ICAR - CNRS - INRP, 24-26 juin 2010.

Malkoun, L., Tiberghien, A. (2008). Objets de savoir et processus scientifiques en jeu dans les productions discursives en classe de physique de lycée. In L. Filliettaz *et al*. (dir.), *Processus interactionnels et situations éducatives*. Bruxelles : De Boeck Supérieur. 67-88

Nonnon, E. (2018). De l'usage des termes construction et coconstruction dans l'analyse des interactions verbales. *Pratiques* 177-178.

Schneeberger, P., Vérin A. (2009). *Développer des pratiques d'oral et d'écrit en sciences. Quels enjeux pour les apprentissages à l'école ?* Paris : INRP.

Corpus multimodal des apprenants en EMILE[4] : constitution, traitements, outils.

Evgenia Bakaldina-Nicol [1]
[1] Laboratoire LLSETI, Université Savoie Mont-Blanc
evgenia.bakaldina@etu.univ-savoie.fr

Introduction

Depuis les années 2000 les logiciels de traitement automatique de langue (TAL)[5] sont de plus en plus impliqués dans la recherche en linguistique de corpus (voir Tutin et al., s.d ; Rohlfing, 2006), ce qui permet notamment d'étudier l'interlangue[6] (Selinker, 1972) au sein des corpus des apprenants (ex. corpus PAROLE[7], le projet Diderot-Longdale[8]). La contribution actuelle traite des défis méthodologiques lors de la constitution (transcription, standardisation) et de codage (pendant et après la transcription) du corpus multimodal des apprenants EMILE.

Dans la première partie nous considérons les questions méthodologiques liées à la transcription. De nombreux chercheurs (Granger et al., 2015 ; Benazzo et Watorek, 2021) soulignent la difficulté de la transcription des données orales, le processus étant très chronophage. Au-delà du choix de type de transcription - **manuelle ou assistée -** le chercheur est confronté aux défis d'exhaustivité de transcription (que transcrire ?) ou de segmentation du **flux de parole en unités de sens (comment transcrire ?)**. **Il est indéniable que l**es choix faits pendant la transcription influent le calcul du nombre de mots dans le corpus et par conséquent ont l'incidence sur les analyses effectuées ultérieurement. Comment concilier les contraintes de transcription et les exigences des **conventions de transcription, comme CHILDES**[9] ?

Dans la seconde partie nous traitons les défis méthodologiques liées au codage. L'interprétation du discours des apprenants peut être délicate lorsqu'il s'agit de catégoriser une erreur (voir Benazzo et Watorek, 2021). Quel codage attribuer aux mots ayant un sens multiple (ex. « right ») ou réalisés dans une graphie non-normalisée (ex. « chose ») ? Comment trouver un compromis entre le codage exhaustif et les contraintes des analyses

[4] L'Enseignement d'une Matière par Intégration d'une Langue Etrangère

[5] Le Traitement Automatique des Langues (TAL) est un domaine scientifique pluridisciplinaire qui se situe au croisement de la linguistique et de l'informatique, souvent associé à l'intelligence artificielle.

[6] 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. (Selinker, 1972 ; Quivy et Tardieu, 2002).

[7] https://slabank.talkbank.org/access/English/PAROLE.html

[8] https://www.clillac-arp.univ-paris-diderot.fr/projets/longdale, consulté le 17/05/2023

[9] CHild Language Data Exchange System - système doté d'outils pour l'analyse des interactions discursives et qui sert comme convention de codage des corpus reconnue un niveau mondial (MacWhinney, 2000).

quantitatives (la comptabilisation d'une erreur plusieurs fois), d'autant plus que le calcul des mots est différent en fonction du logiciel TAL utilisé[10]. Nous illustrons des avantages et des inconvénients des outils TAL (CLAN et EXMARaLDA pour les productions orales, UAM CorpusTool et AntConc pour les productions écrites) avec des exemples tirés du corpus multimodal EMILE (voir le contexte ci-dessous). La présentation se termine avec une conclusion comprenant les éléments à prendre en compte lors de la constitution du corpus.

Contexte de la recherche et méthodologie

La contribution se situe dans le cadre plus large d'un projet d'observation d'une classe EMILE sur un an dans un contexte de lycée français (étude de cas). 16 heures de cours (histoire-géographie en L2 : anglais) ont été enregistrés ; 280 documents (supports des professeurs et productions des élèves) ont été récoltés. Les données ont été transcrites, le corpus multimodal (écrit+oral) de 119.000 mots a été constitué et analysé. Le fonctionnement de l'interlangue des élèves, ainsi que l'interaction entre l'input et l'output sont étudiés par le biais de l'observation des cours EMILE. Les composantes linguistiques du corpus sont analysées grâce aux logiciels : Exmaralda, CLAN, Hyperbase, SketchEngine, UAM Corpus Tool, AntConc. Le travail de recherche étudie dans quelle mesure un statut particulier de la langue en cours EMILE (à la fois un objet et un outil d'apprentissage) modifie les conditions et les résultats de l'apprentissage.

Constitution du corpus multimodal

La constitution d'un corpus soulève des questions relatives à la transcription, la standardisation et le codage des erreurs pendant et après la transcription[11]. Le codage des productions des élèves prend en compte des disfluences principales, des erreurs de lexique, de grammaire et de syntaxe. Tableau 1 présente de différentes étapes de constitution du corpus EMILE ainsi que des défis méthodologiques principaux lors du traitement du corpus dans EXMARaLDA.

Constitutio	Transcription des données	Codage pendant la transcription	Nettoyage et standardisation	Codage après la transcription

[10] Le logiciel SketchEngine calcule les lemmes, tandis que AntConc compte le nombre des mots.

[11] Par exemple, selon la convention CHILDES il faut écrire les mots erronés dans une orthographe normalisée afin que les mots soient pris en compte lors de l'analyse du corpus.

n d u c o r p u s	• Enregistrements, documents, productions écrites des élèves • Phénomènes méta- et supralinguistiques (toux, gestes, rires) ne sont pas pris en compte • Contenu textuel est gardé uniquement	Productions des élèves : • Corpus oral : disfluences principales. • Corpus écrit : orthographe. • Corpus écrit et oral : lexique, grammaire, syntaxe.	Semi-automatique : • Chiffres en lettres, • Abréviations acceptées internationalement • Pas de liens, sites Internet.	• Erreurs et corrections sont gardées : *not to <invaded*> invade Cuba* • Certaines données cachées : -mots en L1 <*guerre*>, -données non pertinentes *<I don't know how to say>*
D é f i s	• Processus chronophage • Que transcrire ? • Qualité des enregistrements et du son, segmentation des phrases ont un impact sur le calcul des erreurs • Vérification manuelle s'impose	Codage dans les deux couches impacte le calcul du nombre d'erreurs : *the politic* situation* *1) politic*0* (mot non-terminé) 2) $MOR $SUF $LOS	Vérification manuelle s'impose : 1950s devient *nineteen fiftyS* au lieu de *nineteen fiftIES*	Comment coder les mots en grammaire non normée ? *Gonna, yea,*etc. Erreurs+corrections=> incidence sur la taille du corpus

Etapes de constitution du corpus et défis méthodologiques.

Segmentation des énoncés

La transcription dans CLAN doit obéir à certaines règles et un protocole précis. Par exemple, les données métalinguistiques doivent se trouver dans un ordre précis ; la ponctuation est obligatoire à la fin de l'énoncé, en revanche, pas de ponctuation possible à l'intérieur de l'énoncé ; les caractères minuscules sont obligatoires, même en début de l'énoncé, sinon CLAN traite ces mots en tant que noms propres. Prenons un exemple.

LV: *It's one of the five giants, remember ?*

Du fait que la ponctuation n'est pas possible à l'intérieur du segment, celui-ci risque d'être classé par le logiciel comme une question puisqu'il y a un point d'interrogation à la fin. Cependant, les indices prosodiques indiquent qu'il s'agit d'une affirmation suivie par *remember* prononcé avec un ton interrogatif qui suggère une question. Ainsi, la solution est de scinder ce segment en deux pour en faire deux énoncés distincts (affirmatif + interrogatif).

LV: *it's one of the five giants. remember ?*

Le principal intérêt de la ponctuation (le point (.)) dans CLAN réside en la délimitation d'un énoncé (*utterance boundary*), ce qui va déterminer notamment le calcul de la longueur moyenne des énoncés (*mean length of utterance, MLU*). Décider comment découper le discours est donc important. Il n'existe pas un seul critère de délimitation des énoncés ; il faut souvent associer critères syntaxiques (et parfois sémantiques), prosodiques et pausologiques. Ce sont généralement les conjonctions de coordination qui posent le plus de problèmes. Dans la grammaire traditionnelle, des propositions liées par *and*, *but* ou *so* constituent une seule phrase, mais à l'oral il est possible d'avoir tout un récit constitué de propositions liées par des

conjonctions. Faut-il alors le considérer comme un seul énoncé ? Cela pourrait être une solution, mais fausserait sans doute les calculs.

L'exemple suivant représente un énoncé ayant des fonctions syntaxiques multiples avec ou sans dépendance de multiples propositions à l'intérieur de ce même segment.

(1) *# er after the war (.) many people # er [/] people had learnt to (*) ((1,1s)) # er work together without caring about (*) (.) social classes or genders or all these things*
(2) ***so** they wanted the government to ((1,7s)) go deeper in their life **and** to ((1,0s)) have more impact on (.) health <or or or> [/] or work, or <like these things> (*)*
(3) ***and so** # er it's important because it's the first time that the Labour Party is (*) elected **and** it will be able to (.) [/] to answer to <what they> [//] # er what they wanted.*

Nous avons fait le choix de combiner deux stratégies (la prosodie et des indices morpho-syntaxiques/lexicaux) afin de délimiter les énoncés. Il est tout de même utile de délimiter les énoncés à l'oral (afin d'obtenir les calculs comme MLU), désignés dans la transcription comme « espace+point » pour marquer la fin de l'énoncé. Il est à noter cependant que l'une des grandes différences entre CHAT[12] et EXMARaLDA est que chaque nouvel énoncé se transcrit sur une nouvelle ligne dans CHAT mais pas dans EXMARaLDA, qui n'utilise qu'une ligne de transcription par locuteur, où sont indiquées les frontières d'énoncé et de segment (=un tour de parole). Globalement il s'agit probablement de trois énoncés, liés par des connecteurs.

Codage des erreurs

Outre la difficulté liée à la délimitation des segments, se pose la question de la catégorisation des erreurs. Comment coder les mots ayant un sens multiple ? Le mot « right » (=all right) peut avoir des étiquetages différents selon le rôle qu'il joue au sein d'un énoncé :

Tout d'abord il peut avoir la fonction d'un commentaire métalinguistique, d'un filler ou d'un jugement qui porte sur les choix linguistiques. Dans ce cas, il est marqué avec ["] :

(4) *don't forget ignorance **right** ["] because it's about the (*) children and it's important.*

(5) ***<okay** never mind> ["].*[13]

Dans certains énoncés *right* n'est pas un commentaire, mais désigne « tag question » :

(6) *the elections take place **right** ?** (= don't they?) Aucun étiquetage particulier n'est nécessaire ; en revanche *right* confère à un énoncé le statut d'une question.

Right peut aussi contribuer à exprimer une demande[14] ou valider une affirmation juste avant :

(7) *you're allowed to have key words. **okay**? (=please use key words).*

---

[12] CHAT et CLAN font partie du CHILDES. Les conventions CHAT permettent la transcription des fichiers sonores grâce aux règles pré-établies connues sous le nom "CHAT format". CLAN est un logiciel d'analyse des données – les transcriptions effectuées en conformité avec le format CHAT (Ratner and Brundage, 2016, p. 2).

[13] Le corpus contient d'autres mots qui remplissent la même fonction d'appréciation : nice, exactly, yes.

[14] A l'instar des expressions : would you? won't you ?

La difficulté majeure survient lorsque nous rencontrons les cas où l'attribution d'une catégorie précise à une déviance parait difficile. En fonction de la façon d'analyser un mot il peut être défini comme erreur ou pas, voire entrer en tant qu'une erreur dans plusieurs catégories/sous-catégories du schéma de codage. Prenons un exemple tiré des productions écrites des élèves.

(8) *democracy: people **chose** their representative*

Nous rentrons ici dans une interprétation de l'intention de l'élève lorsqu'il a écrit *chose.* La prudence s'impose car le corpus doit être représentatif (= le plus fidèle) du discours des apprenants sans le sous- ou surinterpréter.

Deux questions se posent. Premièrement, quel temps grammatical était visé par l'élève, le passé ou le présent ? Deuxièmement, s'il y a une erreur, s'agit-il d'une déviance purement morphologique (confusion entre les formes du présent et du passé) ou est-ce que la déviance a des origines phonétiques ? A savoir, l'élève à l'intention d'écrire le présent de *choose*, mais ne maîtrise pas la prononciation du verbe et emploie une orthographe phonétique. Parlons-nous d'une maîtrise complète ou partielle du verbe ?

En fonction des réponses à cette série de questions *chose* peut figurer dans :

- aucune catégorie (pas d'erreur) ;
- la sous-catégorie **erreur/morphologie/temps ;**
- la sous-catégorie **erreur /orthographe/orthog-verbe.**

N'ayant pas de possibilité d'interroger l'élève par rapport à ces intentions, nous devons recourir aux stratégies alternatives d'analyse. La première question trouve sa réponse en observant l'environnement co-textuel proche de celui dans lequel *chose* est employé afin d'observer la systématicité de son emploi. Les observations montrent que l'emploi de *choose* est erroné dans 3 cas sur 4, nous supposons que l'erreur est plutôt systématique, ce qui exclut une hypothèse de lapsus ou juste « *a slip of the pen* ». Ainsi, la déviance en question est susceptible d'être classée soit comme une erreur (méconnaissance des règles grammaticales) soit comme une faute (déviance corrigible par l'apprenant grâce à l'indication venant de l'extérieur). Quel que soit le statut, nous pouvons désormais le coder dans UAM Corpus Tool de façon plus générale : **erreur/morphologie/verbe**. Cependant, la question plus délicate de l'origine de l'erreur (grammaticale ou phonétique) reste irrésolue.

Par ailleurs, nous sommes confrontés au conflit entre l'annotation exhaustive (la plus complète possible) et les analyses quantitatives. Par exemple, l'erreur morphologique peut être comptée plusieurs fois suite à une annotation multiple : l'erreur d'accord, l'erreur d'ajout, l'erreur dans la partie du discours. Comment faire pour trouver un compromis raisonnable ? Plusieurs solutions seraient envisageables : compter non pas la description des erreurs mais ce qui signale l'erreur (par exemple, *), puisque ce marqueur n'apparaît qu'une seule fois ; avoir un codage générique que l'on noterait dans l'annotation avec un code indiquant que ceci est une erreur ($ERR suivi d'une description détaillée). Cela comporte le risque d'être redondant à chaque fois. Par conséquent, la première solution est tout de même préférable si nous souhaitons compter le nombre d'erreurs.

Finalement, quels outils de traitement de données choisir ? Il est nécessaire de prendre en compte les spécificités des logiciels en les rapportant aux objectifs de recherche. AntConc et Sketch Engine disposent des fonctionnalités similaires, mais comptent les mots différemment.

SketchEngine peut être configuré pour calculer le nombre de lemmes, mesure plus précise pour certaines recherches (ex. la fréquence des mots). AntConc est plus pertinent pour le traitement des corpus oraux, puisqu'il est configurable afin de ne pas prendre en compte les informations placées entre les balises (< >). Cependant, pour être traité, le texte doit être en format TXT (raw text), sinon en HTML ou XML ; l'exploration au format HTML ou XML est limitée, car des symboles contenus dans les titres (ex.<H1> à <H6>), ne sont pas pris en compte par AntConc.

Conclusions

La compilation du corpus doit obéir aux règles suivantes : être représentative du langage des apprenants (= représenter des phénomènes linguistiques ; Rastier, 2004) ; correspondre aux normes langagières et aux exigences des conventions communes (comme CHILDES) ; être systématique et régulière (codage) afin de permettre des prolongements possibles d'exploration (Granger, 2015).

Une transcription est censée, d'un côté, représenter les phénomènes linguistiques observés dans les énoncés. D'un autre côté, toute représentation du texte sous une forme graphique est soumise à des normes précises, autrement dit, les données doivent être transcrites conformément aux conventions adoptées internationalement. Ceci est atteignable au moyen des codes de transcription.

Il n'existe pas un seul critère de délimitation des énoncés ; il faut souvent associer critères syntaxiques (et parfois sémantiques), prosodiques et pausologiques. Les pauses (ou leur absence), peuvent être des indices précieux dans le découpage des segments mais ne sont pas toujours des marqueurs fiables de ponctuation dans tous les cas.

Ainsi, il est important de trouver un compromis raisonnable entre des exigences spécifiques des logiciels TAL et des conventions communément employées, tout en prenant en compte les avantages et les limites des outils de transcription et d'analyse des données.

Références bibliographiques

Bakaldina-Nicol, E. (2023). *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].

Benazzo, S. et Watorek, M. (2021). Transcription de corpus oraux d'apprenants débutants en français L2 : quelques enjeux théoriques. Dans : Lorenzo Spreafico, Giuliano Bernini, Ada Valentini & Jacopo Saturno (éds.) *Superare l'evanescenza del parlato. Un vademecum per il trattamento digitale di dati linguistici* (pp. 127-165). Bergamo: Sestante, pp.127-165, 2021. ffhal-03312984f

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

Biber, D. (2012). Corpus-based and corpus-driven analyses of language variation and use. Dans B. Heine et H. Narrog (dir.), *The Oxford handbook of linguistic analysis* (p. 160-191). Oxford University Press.

Doyle, W. (1986). Classroom organization and management. Dans Wittrock, M. (dir.). *Handbook of research on teaching* (3rd ed.), (p. 392-431). Macmillan.

Granger, S., Meunier, F. et Gilquin, G. (dir.) (2015). *Cambridge Handbook of Learner Corpus Research.* Cambridge University Press.

James, C. (1998). *Errors in language learning and use*. *Exploring error analysis*. Longman.

MacWhinney. (2000). *The CHILDES Project Tools for Analyzing Talk – Electronic Edition Volume 1 : Transcription Format and Programs. Part 1: The CHAT Transcription Format*. Lawrence Erlbaum Associates.

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

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

Ratner, B.N. et Brundage, S. (2016). *A clinician's complete guide to CLAN and PRAAT.* [Manuel en ligne. George Washington University].

URL: https://vandammark.com/WSU/BernsteinRatnerBrundage_2016_ClinClan.pdf

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. Dans: *Gersprachsforschung* [Workshop report; en ligne]. (7), 99-123.

URL: http://www.gespraechsforschung-ozs.de/heft2006/tb-rohlfing.pdf

Schmidt, T. (2010). *Exmaralda. EXAKT 1.0.* [Manuel en ligne]. URL:http://www.exmaralda.org/files/EXAKT_Manual.pdf

Schmidt, T. et *Wörner,* K. (2011). Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics,* 19 (4).

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*. 10 (3), 219-31.

Sinclair, J. et Coulthard, M. (1975). *Towards an analysis of discourse*. Oxford University Press.

Tognini-Bonelli, E. (2001). Corpus Linguistics at work. *Studies in corpus linguistics*. John Benjamins Publishing.

Tutin, A., Jaques, M-P., Kraif, O. et Hartwell, L. (s.d.). Introduction à la linguistique de corpus. Université Grenoble Rhône-Alpes [Cours en ligne de l'UGA].

Sitographie

Anthony, L. (n.d.). *Laurence Anthony's Homepage*. https://www.laurenceanthony.net/

Exploration de corpus : outils et pratiques. (n.d.). http://explorationdecorpus.corpusecrits.huma-num.fr/

Schmidt, T. (2023, January 15). *EXMARaLDA*. https://www.exmaralda.org/

UAM CorpusTool Homepage. (n.d.). http://www.corpustool.com/

Outiller l'étude des chaînes de référence dans des écrits scolaires

Martina Barletta[1], Claude Ponton[1]
[1] Laboratoire LIDILEM, Université Grenoble Alpes
martina.barletta@univ-grenoble-alpes.fr, claude.ponton@univ-grenoble-alpes.fr

La recherche Scolinter (Ponton *et al.*, 2021) s'intéresse à l'étude des compétences en écriture des élèves de primaire en France, en Italie et en Espagne. La communication actuelle cible plus particulièrement les compétences relevant de la textualité qui contrairement aux compétences orthographiques ont fait l'objet de moins d'études du point de vue de la production et en particulier chez les enfants (Bonnemaison, 2018). Parmi les facteurs qui réalisent la textualité, notre recherche se concentre actuellement sur la question des chaînes de référence. Il s'agit pour nous de décrire linguistiquement comment les élèves gèrent cet aspect dans leur production d'écrits, comment cet élément reflète les compétences de construction de la cohérence et de la cohésion textuelles, comment ces compétences évoluent dans le temps et quelles sont les similarités et les différences entre les trois langues.

Depuis 2018, nous constituons un large corpus longitudinal d'écrits d'élèves comparables dans ces trois langues (Ponton *et al.*, 2021), le corpus éponyme *Scolinter*[15]. A terme, ce corpus sera composé d'un peu moins de 7.000 textes ce qui rend complexe et coûteux une exploitation uniquement manuelle. Si diverses études (Delaborde, 2020 ; Grobol, 2020 ; Landragin, 2016 ; Muzerelle *et al.*, 2013 ; Wilkens *et al.*, 2020) se sont intéressées au traitement automatique de la coréférence dans des textes d'adultes, les outils qui parfois en découlent ne sont pas toujours accessibles ou maintenus et ne semblent pas forcément adaptés à nos objectifs actuels. Ainsi, pour assister les chercheurs dans la description de la coréférence sur des textes d'élèves dans les trois langues, nous cherchons à développer une méthodologie et un outillage permettant une pré-annotation des chaînes de référence (Barletta, 2022).

Nous considérons les chaînes de référence comme l'ensemble d'au moins trois expressions faisant référence à la même entité, appartenant à l'univers du texte (Schnedecker, 1997). Si on a moins de trois composantes, les notions d'anaphore et de coréférence sont suffisantes pour une description pertinente des phénomènes de paires d'enchaînements référentiels. Cependant, nous pouvons faire l'hypothèse que la compétence des auteurs des textes analysés puisse aussi conditionner les modalités de construction de ces chaînes et le type de mentions qu'on peut y retrouver plus souvent, dans le sens où l'on va retrouver dans les productions des élèves des structures plus simples que dans les écrits des scripteurs experts et une variété lexicale assez réduite dans les mentions utilisées.

La première étape de cette recherche, qui fait l'objet de cette communication, consiste à proposer une méthodologie pour arriver à terme à un modèle d'annotation répondant à ces objectifs. Cette communication fera tout d'abord un état des recherches sur l'annotation des chaînes de référence et de la coréférence avant de présenter les spécificités du corpus *Scolinter* et enfin d'exposer l'état actuel du travail.

[15] Disponible sur son site dédié http://scoledit.org/scolinter/

Travaux et corpus antérieurs

La linguistique de corpus française s'intéresse depuis des décennies à l'étude et la description des phénomènes de construction de la textualité comme l'anaphore et les chaines de référence (Charolles, 1988, p. 198 ; Chastain, 1975 ; Corblin, 1985, 1995 ; Schnedecker, 1997). Bien que des corpus annotés existent déjà en français au début des années 2000, leurs caractéristiques ne les rendent pas globalement représentatifs de la coréférence ou utilisables pour l'apprentissage profond, soit à cause des limitations dans les types d'anaphore codés (Tutin *et al.*, 2000), soit à cause de la taille du corpus (Gardent *et al.*, 2005). Cependant, ces projets ont contribué à ouvrir la voie aux réflexions sur l'annotation de ces phénomènes sur des corpus de grande taille comme par exemple *Annodis* (Péry-Woodley *et al.*, 2011), *ANCOR* (Garcia-Debanc *et al.*, 2019) pour l'oral spontané, et *DEMOCRAT* (Landragin, 2016) pour la langue écrite. Dans notre domaine d'étude des écrits scolaires, il faut noter l'existence du corpus *RésolCo* (Garcia-Debanc *et al.*, 2021, 2019) qui représente le seul corpus français de taille moyenne annoté en continuité référentielle sur ce type d'écrits. Les trois projets *ANCOR*, *DEMOCRAT* et *RésolCo* seront décrits plus précisément dans la communication car ils constituent les références de notre propre recherche.

Le corpus *ANCOR* représente « le premier corpus d'oral spontané d'envergure annoté en coréférence » pour la langue française et « distribué librement » (Muzerelle *et al.*, 2013). Il est composé de plusieurs corpus de parole spontanée transcrite. Son objectif était de répondre au manque de corpus pour entrainer un système de résolution de la coréférence en français (Muzerelle *et al.*, 2013), dans un moment où d'autres langues majoritaires étaient déjà dotées de tels corpus. Cependant, si ce corpus était représentatif de la langue parlée, un corpus écrit dans lequel toutes les expressions référentielles sont annotées était encore absent du panorama. Le corpus *DEMOCRAT* répond à ce manque. Il constitue « le premier corpus de grande taille librement disponible pour le français écrit » (Landragin, 2021, p. 12). L'objectif du projet était de constituer un corpus diachronique de textes, écrits entre le 12e et le 21e siècle, relevant de genres textuels variés, « et d'en autoriser des exploitations par des outils de traitement automatique des langues (TAL), plus précisément par des outils faisant appel à de l'apprentissage profond » (Landragin, 2022, p. 49). Un des choix méthodologiques du projet *DEMOCRAT*, qui le démarque de projets similaires, est le fait d'avoir annoté pour la première fois en France les expressions référentielles sur l'intégralité des textes du corpus. (Landragin, 2021, p. 12).

Dans le domaine des corpus d'écrits scolaires, le corpus *RésolCo* (Garcia-Debanc *et al.*, 2017, 2021) a abordé l'annotation de la continuité référentielle sur des écrits de niveaux scolaires variés. Il a été récolté dans des classes de niveaux différents, du CE2 à l'université (Garcia-Debanc *et al.*, 2021). En s'appuyant sur l'expérience d'annotation faite lors de la conception d'*Annodis* (Péry-Woodley *et al.*, 2011) ainsi qu'au manuel d'annotation du corpus *Democrat*, le corpus *RésolCo* se pose le double objectif de constituer une ressource annotée en continuité référentielle et d'élaborer une cartographie des formes linguistiques qui manifestent les compétences textuelles et discursives en cours de développement (Garcia-Debanc *et al.*, 2021). La plupart des textes qui constituent ce corpus ont été produits à partir de la même consigne imposant aux élèves la résolution de problèmes de cohésion textuelle (Garcia-Debanc & Bonnemaison, 2014 ; Garcia-Debanc & Bras, 2016).

Ces différents corpus, tout en ayant en commun un intérêt pour la manifestation des phénomènes de cohérence et de cohésion textuelles, présentent des spécificités, à la fois liés aux genres textuels annotés et aux objectifs que les annotations contribuent à réaliser.

Le corpus Scolinter

Le projet *Scolinter*[16] (Ponton *et al.*, 2021) cherche à étudier les compétences en littéracie des élèves à chaque niveau, ainsi que l'évolution de ces compétences dans les trois langues tout au long de l'école primaire. Pour atteindre cet objectif, les chercheurs du projet développent le corpus *Scolinter* qui comporte des textes d'élèves recueillis dans les trois pays à partir de la même consigne. Ce corpus devrait être finalisé d'ici 2025. Actuellement, il est composé de la partie longitudinale du corpus Scoledit (Wolfarth *et al.*, 2018) pour le français (CP à CM2) et des textes de CP, CE1 et C2 pour l'espagnol et l'italien ; les textes de CM1 et CM2 (respectivement grades 4 et 5) sont en cours de recueil et de transcription dans ces deux pays. A terme, le corpus sera constitué de 1824 textes français, 3250 textes italiens et 1910 textes espagnols. Pour permettre le recours à des outils de traitement automatique, le corpus propose une version normalisée de chacun des textes. A noter que cette normalisation, réalisée manuellement, ne rectifie principalement que les erreurs de type orthographique. C'est sur cette version normalisée que porte notre travail autour de l'annotation outillée des chaînes de référence. Toutefois, de par leur faible longueur et leur consigne spécifique, les textes de CP ne seront pas pris en compte dans notre étude.

Vers un outil d'annotation des chaînes de référence

Comme énoncé précédemment, l'un de nos objectifs est d'outiller l'annotation des chaînes de référence dans les écrits scolaires dans les trois langues pour faciliter la description linguistique des compétences autour de la textualité. Le cadre théorique auquel on fait référence dans la définition de ce phénomène est celui largement partagé à l'heure actuelle dans le domaine (Corblin, 1985 ; Schnedecker, 1997, 2005 ; Charolles, 2002).

Ce travail, effectué sur un corpus longitudinal est à l'heure inédit et pourrait constituer l'une des premières descriptions du développement de ces compétences à l'écrit pour des enfants de l'école primaire. Suite à une première étude (Barletta, 2022) et dans le même esprit que l'approche RésolCo[17], nous avons décidé de cibler uniquement les chaines de référence liées aux quatre personnages principaux proposés par la consigne à savoir : le chat, la sorcière, le loup et le robot. En effet, ce choix nous assure que tous les textes impliquent au moins l'un de ces personnages ce qui facilite leur comparaison au niveau longitudinal sur une même langue et au niveau contrastif entre les trois langues. Notre modèle d'annotation s'inspire largement du projet RésolCo ce qui nous permettra à terme d'obtenir des résultats comparables avec

[16] Disponible sur son site dédié http://scoledit.org/scolinter/

[17] Dans l'approche RésolCo, « les productions sont issues d'une tâche de résolution de problème de cohésion textuelle, la même pour tous les niveaux scolaires. Trois phrases contenant des pronoms personnels (ils et elle) et des syntagmes nominaux introduits par un déterminant démonstratif (« cette maison », « ce grand bruit », « cette aventure ») sont proposées aux étudiants. […] » (Barletta, 2022, p. 36) Les textes sont « annotés selon les référents principaux de la tâche (à savoir les référents reliés par le scripteur aux pronoms personnels et au syntagme nominal pluriel explicités dans les trois phrases qui constituent la consigne : « elle », « il » et « les enfants », indiqués en gras dans les phrases). » (*ibid.*)

ceux obtenus par ce projet (Garcia-Debanc et al., 2021). Notons ici que la partie française du corpus Scolinter, comme le corpus RésolCo, font partie du projet E-Calm dont le but était de structurer et mettre à disposition de la communauté scientifique des corpus d'écrits d'élèves et d'étudiants enrichis d'annotations linguistiques sur des différents phénomènes et niveaux de langage. (Guide d'annotation RésolCo, 2022)

Lors de cette première étude, un outil d'annotation a été développé sur un modèle simple, dont l'architecture a été décrite par M. Barletta (2022). Ces premiers résultats servent de base au modèle plus complet en cours de définition. Pour le moment, en nous appuyant sur les projets DEMOCRAT et RésolCo, nous avons choisi d'annoter :

- les syntagmes nominaux,
- les pronoms personnels,
- les verbes en l'absence d'un sujet explicité
- les déterminants possessifs.

Nous avons également décidé d'annoter la marque anaphorique sur les verbes, pour garantir la possibilité de réaliser des analyses homogènes avec l'italien et l'espagnol, où le sujet peut être souvent omis.

Notre communication proposera la présentation des premiers résultats obtenus en appliquant cette première version de notre modèle d'annotation sur les textes du corpus de niveau CE2. La suite du travail consistera à développer un corpus de référence qui nous permettra de tester et d'affiner ce modèle. Enfin, nous prévoyons d'évaluer les outils d'annotation existants pour les adapter à notre problématique.

Références bibliographiques

Barletta, M. (2022). *DeCorScol : Conception d'un outil d'assistance à l'annotation des chaînes de coréférence dans les écrits scolaires* [Mémoire de master 2, Université Grenoble Alpes]. https://dumas.ccsd.cnrs.fr/dumas-03826763v1/

Bonnemaison, K. P. (2018). *Anaphore et référence en production écrite : Étude de textes narratifs d'élèves de 9 à 11 ans, du CE2 au CM2* [Thèse de doctorat, Université Toulouse le Mirail - Toulouse II]. https://tel.archives-ouvertes.fr/tel-02627042

Charolles, M. (1988). Les plans d'organisation textuelle : Périodes, chaînes, portées et séquences. *Pratiques*, *57*(1), 3-13. https://doi.org/10.3406/prati.1988.1468

Charolles, M. (2002). *La référence et les expressions référentielles en français* (Ophrys).

Chastain, C. (1975). Reference and Context. *Language, Mind, and Knowledge*, *7*, 194-269.

Corblin, F. (1985). Les chaînes de référence : Analyse linguistique et traitement automatique. *Intellectica*, *1*(1), 123-143. https://doi.org/10.3406/intel.1985.851

Corblin, F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes. https://jeannicod.ccsd.cnrs.fr/ijn_00550962

Delaborde, M. (2020). *Analyse en corpus de chaînes de coréférence : La coréférence non-stricte à l'épreuve de la linguistique outillée* [Thèse de doctorat, Université de la Sorbonne nouvelle - Paris III]. https://theses.hal.science/tel-03425446

Garcia-Debanc, C., & Bonnemaison, K. (2014). La gestion de la cohésion textuelle par des élèves de 11-12 ans : Réussites et difficultés. *SHS Web of Conferences*, *8*, 961-976. https://doi.org/10.1051/shsconf/20140801349

Garcia-Debanc, C., & Bras, M. (2016). *Vers une cartographie des compétences de cohérence et de cohésion textuelle dans une tâche-problème de production écrite réalisée par des élèves de 9 -12 ans : Indicateurs de maitrise et progressivité*. *Recherches textuelles*(13). https://hal.science/hal-01987031

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., & Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, (16). https://doi.org/10.4000/corpus.2783

Garcia-Debanc, C., Ho-Dac, L.-M., Federzoni, S., Bras, M., & Rebeyrolle, J. (2019, novembre 6). *ResolCo un corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence*. 10èmes Journées Internationale de la Linguistique de Corpus. https://hal.science/hal-02877122

Garcia-Debanc, C., Rebeyrolle, J., & Ho-Dac, L.-M. (2021). La continuité référentielle dans le corpus RÉSOLCO : Méthode d'annotation et premières analyses. *Langue francaise*, *211*(3), 99-114.

Gardent, C., & Manuélian, H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *Revue TAL*, *46*(1), 115.

Grobol, L. (2020). *Coreference resolution for spoken French* [Thèse de doctorat, Université Sorbonne Nouvelle - Paris 3]. https://hal.archives-ouvertes.fr/tel-02928209

Guide d'annotation RésolCo. (2022). https://hodaclm.github.io/resolco/

Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, *92*, 11.

Landragin, F. (2021). Le corpus Democrat et son exploitation. Présentation. *Langages*, *224*, 11-24.

Landragin, F. (2022). Expressions référentielles et chaînes de référence en français : Le projet Democrat et son exploration des rapports entre linguistique textuelle et linguistique de corpus. *Écho des études romanes*, *18*(1), 49-65. https://doi.org/10.32725/eer.2022.004

Muzerelle, J., Lefeuvre, A., Antoine, J.-Y., Schang, E., Maurel, D., Villaneau, J., & Eshkol, I. (2013). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In ATALA (Éd.), *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles*, 555-563. https://hal.archives-ouvertes.fr/hal-01016562

Péry-Woodley, M.-P., Afantenos, S., Ho-Dac, L.-M., & Asher, N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Revue TAL*, *52*(3), 71.

Ponton, C., Gutiérrez-Caceres, R., Teruggi, L., Farina, E., Brissaud, C., & Wolfarth, C. (2021). Scolinter : Un corpus trilingue. L'exemple de la segmentation en mots. *Langue française*, *211*(3), 37-50. https://doi.org/10.3917/lf.211.0037

Schnedecker, C. (1997). *Nom propre et chaînes de référence* (Klincksieck, Éd.; Vol. 21). Librairie KLINCKSIECK. https://hal.archives-ouvertes.fr/hal-00808797

Schnedecker, C. (2005). Les chaînes de référence dans les portraits journalistiques : Éléments de description. *Travaux De Linguistique*, *51*. https://doi.org/10.3917/tl.051.0085

Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, É., Zaenen, A., Rayot, S., & Antoniadis, G. (2000). Annotating a large corpus with anaphoric links. *Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC2000)*, 2. https://hal.archives-ouvertes.fr/hal-00373327

Wilkens, R., Oberle, B., Landragin, F., & Todirascu, A. (2020). French Coreference for Spoken and Written Language. *Proceedings of the 12th Language Resources and Evaluation Conference*, 80-89. https://aclanthology.org/2020.lrec-1.10

Wolfarth, C., Brissaud, C., & Claude, P. (2018). Transcrire et normer un corpus scolaire : Pour quelles analyses? In C. Brissaud, M. Dreyfus, & B. Kervyn (Éds.), *Repenser l'écriture et son évaluation au primaire et au secondaire,* 121-145. Presses universitaires de Namur. https://books.openedition.org/pun/5283?lang=en

Disfluencies and directionality in simultaneous interpreting. A corpus study comparing into-B and into-A interpretations from the European Parliament

Magdalena Bartłomiejczyk,[1, 2] Ewa Gumul[1]
[1] Univeristy of Silesia, Poland
[2] University of Vienna
magdalena.bartlomiejczyk@us.edu.pl, ewa.gumul@us.edu.pl

Introduction

Simultaneous interpreters have to manage and coordinate concurrent source language comprehension and target language production, often unable to predict the speaker's ultimate communicative intent. This produces significant cognitive load, which, however, normally fluctuates across an interpreting task. Research has revealed symptoms of increased cognitive load, detectable either in the interpretation or in the interpreter's physical reactions (Chen 2017). The former include disfluencies such as filled and silent pauses (Plevoets & Defrancq 2018; Gumul 2021) that lie in the focus of our interest. At the same time, disfluencies can also be perceived as presentation errors (Kurz & Färber 2003; Bartłomiejczyk 2010), i.e., symptoms of deteriorating interpreting quality.

International organizations generally favour interpreting into A (the native language) due to the widespread belief in the superiority of this direction. It tends to be seen as both easier for the interpreter and conducive to better quality. However, into-B interpreting often fulfils a genuine market need, and this is definitely the case for interpreting from languages of lower diffusion (i.e., rarely learned by non-natives), including Polish.

A plethora of empirical studies shows directionality effects for various language combinations. In this vein, we would like to explore disfluencies in A-B vs. B/C-A interpreting by comparing output in each of the interpreting directions by European Parliament (EP) interpreters working between Polish and English. Native speakers of English only interpret from Polish into English, while some native speakers of Polish interpret both ways. Our quantitative analysis focuses on three types of disfluencies that have been releatively widely researched in Interpreting Studies: anomalous pauses, hesitation markers, and false starts. If, in accordance with the common belief (e.g., Chmiel 2016), cognitive load is lower for interpreting into A, Polish-English interpretations by Poles (A-B) should exhibit significantly more disfluencies of various types than both English-Polish interpretations by Poles (B/C-A) and Polish-English interpretations by Brits (C-A): Hypothesis 1. Furthermore, interpretations into the native language, be it Polish or English, should feature comparable prevalence of disfluencies: Hypothesis 2. Finally, individual bidirectional interpreters should

produce more disfluencies when working into their B than when working into A: Hypothesis 3.

Our project has considerable practical importance for the setting under investigation, i.e., EU institutions. It is meant to provide hard data on one aspect of directionality that, among others, may guide the general human resources policies of the EU interpreting services and influence their outlook on into-B interpreting. Excessive disfluencies lower the quality of the interpreting product, particularly from the perspective of the audience (e.g., Pradas Macías 2006). If into-B interpreting is actually shown to produce target texts of markedly inferior quality, the need for it might be substantially reduced over time through appropriate recruitment and professional development policies. If, however, the existing concerns over the quality of into-B interpreting are ungrounded, maybe there is no need to foster into-A interpreting to the extent it is fostered today.

Corpus and methodology

Corpus

The material for our analysis has been extracted from EP-Poland (see Bartłomiejczyk et al. 2022), a large bidirectional parallel corpus containing all Polish and English contributions to eleven plenary debates of the EP. The debates, held in the years 2016-2020, were selected because of their topic, i.e., the current developments in Poland related to the rule of law crisis and the resulting conflict of the Polish government with the EU. The main aim when compiling the corpus was to obtain material for discourse analytic explorations focusing on ideology, however, the topic also ensured relatively frequent use of Polish as a source language. Consequently, the share of English-Polish and Polish-English interpretations is fairly balanced across the corpus. While interpretations into Polish are exclusively provided by native speakers of Polish, interpretations into English are provided by both native and non-native speakers of English. For the needs of this analysis, three subcorpora were necessary: PL-A with interpretations from English into Polish by native speakers of Polish, EN-A with interpretations from Polish into English by native speakers of English, and EN-B with interpretations from Polish into English by native speakers of Polish.

Methodology

Individual interpreters were identified in the recordings semi-automatically by the timbre of their voice using the X-vector method (see Bartłomiejczyk et al. 2022 for details). Within the whole corpus, 36 individuals were identified, out of whom ten interpret in both directions. However, a threshold for inclusion into the analysis needed to be established so as to account for variable source language input. Considering that the EP speeches are very short (2 minutes 16 seconds on average across EP-Poland) and that disfluencies are likely to occur frequently, we settled on five minutes and at least two different speeches for each interpreting direction. Six bidirectional interpreters meet the criteria to be included both in the PL-A and EN-B subcorpora. Apart from them, 15 other interpreters were included in the PL-A subcorpus and three in the EN-B subcorpus, i.e., PL-A contains output from 21 interpreters (5 hours of material), and EN-B from nine interpreters (nearly 2 hours). The big difference in the size of the two subcorpora results from the fact that many Polish interpreters do not work into English and some who do provided too little output to qualify. Out of eleven interpreters in EP-Poland who only work into English, eight exceed the threshold as described above.

However, a phonetic analysis of divergences from native pronunciation norms (Bartłomiejczyk and Rojczyk under review) conclusively shows that only five of them are definitely native speakers of English. Consequently, the output of these five interpreters is included in our EN-A subcorpus (nearly 3 hours). As we are particularly interested in the performance of individual bidirectional interpreters, their output was further extracted to create smaller subcorpora EN-B/1 and PL-A/1. Their length is 1 h 30 min and 1 h 20 min, respectively.

The disfluency phenomena in which we are interested, i.e., hesitation markers, false starts, and anomalous pauses, were annotated already at the stage of manual transcription of the interpretations to be included in the EP-Poland corpus. Each interpretation was transcribed on the basis of the recording retrieved from the EP website and afterwards verified by another person (see Bartłomiejczyk et al. 2022). Hesitation markers, coded uniformly as <@> irrespective of their actual sound, are non-lexical fillers mainly in the form of prolonged vowels. False starts are retraced and non-retraced truncations at the word level, coded with a hyphen following the interrupted word, e.g., <pol->. Finished words, even if followed by a repair, are not treated as false starts in our taxonomy. Anomalous pauses (coded as <--->) comprise only pauses exceeding three seconds, as such a high threshold should unambiguously point to non-strategic interruptions of the interpreter's speech flow, possibly indicating processing problems.

As the subcorpora, as well as individual interpreters' contributions, are of different lengths, the prevalence of each type of disfluency had to be calculated as a normalized frequency per 100 words of target language output.

Results

As for Hypothesis 1, the only statistically significant difference relates to anomalous pauses, which are more prevalent in EN-A than in EN-B. This finding contradicts the expected directionality effect, i.e., potentially shows an advantage of into-B interpreting. However, it may also be attributable to differing norms related to pausing behaviour in the Polish and English booths, acting as separate communities of practice. On the whole, our Hypothesis 1 has not been confirmed. As for Hypothesis 2, again, a statistically significant difference was shown for anomalous pauses, more prevalent in EN-A than PL-A. As in this case we were trying to confirm the similarity of the subcorpora, this finding goes against the intitial expectations. The fact that the overall convergence between EN-A and PL-A is actually smaller than between EN-B and PL-A also speaks against Hypothesis 2. Finally, the comparison between EN-B/1 and PL-A/1, designed to test differences related to interpreting in each direction by the same individuals, failed to show any statistically significant effects, which does not validate Hypothesis 3.

On the whole, we have found no directionality effects related to disfluencies that could render support to a conclusion that Polish-English interpretations by Brits are characterized by higher quality and/or are performed with more ease than those provided by their Polish colleagues; or that Polish interpreters perform better when interpreting into their native language than vice versa. While into-A interpreting from multiple languages may be universally preferred over A-B interpreting for other reasons, such as avoidance of relay, fluency understood as

minimizing the prevalence of the three problem indicators under analysis has not been shown here to depend on interpreting into or out of the interpreter's native language.

References

Bartłomiejczyk, M. (2010). Effects of short intensive practice on interpreter trainees' performance. In D. Gile, G. Hansen & N.K. Pokorn (Eds.), *Why Translation Studies Matters* (pp. 183–194). John Benjamins.

Bartłomiejczyk, M., Gumul, E., & Koržinek, D. (2022). EP-Poland: Building a bilingual parallel corpus for interpreting research. *GEMA Online Journal of Language Studies 22*(1), 110–126.

Bartłomiejczyk, M. & Rojczyk, A. (under review). How native-like do conference interpreters sound in L2?

Chen, S. (2017). The construct of cognitive load in interpreting and its measurement. *Perspectives 25*(4), 640–657.

Chmiel, A. (2016). Directionality and context effects in word translation tasks performed by conference interpreters. *Poznan Studies in Contemporary Linguistics 52*(2): 269–295

Defrancq, B. & Plevoets, K. (2018). Over-uh-load, filled pauses in compounds as a signal of cognitive load. In C. Bendazzoli, M. Russo, & B. Defrancq (Eds.), *Making Way in Corpus-based Interpreting Studies* (pp. 43–64). Springer.

Gumul, E. (2021). Explicitation and cognitive load in simultaneous interpreting: Product- and process-oriented analysis of trainee interpreters' outputs. *Interpreting 23*(1), 45–75. https://doi.org/10.1075/intp.00051.gum

Kurz, I. & Färber, B. (2003). Anticipation in German–English simultaneous interpreting. *Forum 1*(2), 123–150.

Plevoets, K. & Defrancq, B. (2018). The cognitive load of interpreters in the European Parliament: A corpus-based study of predictors for the disfluency *uh(m)*. *Interpreting 20*(1), 1–32.

Pradas Macías, E. M. (2006). Probing quality criteria in simultaneous interpreting: The role of silent pauses in fluency. *Interpreting* 8 (1), 25–43.

Le subjonctif dans les Enquêtes sociolinguistiques à Orléans : de la norme à l'usage

Fatma Ben Barka Messaoudi
Laboratoire EMA, CY Cergy Paris Université
Fatma.messaoudi1@cyu.fr

Introduction

Il y a très souvent un décalage entre l'évolution d'un fait linguistique et celle de son traitement, qui ne change que très rarement, même si le phénomène en question ne cesse de se modifier et de se métamorphoser. En effet, le fonctionnement du subjonctif de points de vue morphologique, syntaxique, sémantique et pragmatique a fait l'objet de plusieurs études linguistiques (Nordahl, 1969, Nolke, 1985, Soutet, 2000...) Dans tous ces travaux, les linguistes ont eu recours soit à la fabrication d'exemples soit à l'emprunt d'exemples écrits, appartenant généralement aux genres littéraire ou journalistique. Face à la rareté des études menées sur des données orales (hormis quelques études sporadiques sur le français du Québec[18]), nous avons tenté de vérifier le maintien des hypothèses classiques formulées sur ce mode verbal dans l'usage réel de la langue.

Corpus et méthodologie

Décrire tous les emplois du subjonctif a nécessité l'exploration de grandes quantités de données. Pour ce faire, nous avons opté pour des énoncés *in vivo* extraits des Enquêtes sociolinguistiques à Orléans (désormais ESLO), l'un des plus grands corpus de français oral, dont la nature nous a permis non seulement de nous approcher du fonctionnement du subjonctif tel qu'il est, dans sa dimension discursive variationniste ; mais aussi d'examiner son éventuelle évolution à quarante ans d'intervalle. Nous avons *de facto* pu bénéficier de la combinaison inédite entre les pistes synchronique et diachronique offertes par le corpus ESLO. Comme le notent Abouda & Skrovec (2018 : 2)

> Il apparaît donc clairement que, dans le domaine de la documentation du français oral hexagonal, le recul diachronique de 40 ans que permet ESLO est fondamentalement novateur.

Corpus

Afin de pouvoir comparer le comportement de subjonctif dans un contexte interactif et séquentiel à son emploi dans une structure phrastique, nous avons choisi trois modules : entretiens, repas et conférences. Le peu de travaux rassemblant une telle diversité interactionnelle dans l'usage du subjonctif nous a incitée à proposer ce nouvel angle d'observation, en traitant les trois situations de communication ensemble.

Nous récapitulons, dans le tableau suivant, les propriétés quantitatives du corpus en termes de durée et ses structures diastratique et diaphasique.

[18] Laurier (1989) et Kastronic (2016).

Sous-corpus ESLO-MDF	ESLO1 / ESLO1 - MDSUB	ESLO2 / ESLO2 - MDSUB	Total
Durée (min)	2675 (44h35min) Entretiens : 2236 Repas : 224 Conférences : 215	2677 (44h37min) Entretiens : 2230 Repas : 236 Conférences : 211	5352 (89h12min) Entretiens : 1690 Repas : 286 Conférences : 305
Nombre de mots	578629	586369	1 164 998
Nombre de locuteurs	37	35	72

table 3. : Les données du corpus

Méthodologie

Cette composition est le résultat d'une opération de sélection des enregistrements d'ESLO1 et d'ESLO2 au cours de laquelle nous avons veillé à assurer le meilleur équilibrage diastratique possible entre et à l'intérieur de chacun d'entre eux, en termes de profil de locuteurs choisis en fonction des variables de sexe, d'âge et de catégorie socioprofessionnelle (CSP).

La méthode d'échantillonnage de nos locuteurs a été fondée sur les principes suivants :

- nous appuyer sur trois CSP : Cadres, Employés, Ouvriers ;
- classer les locuteurs choisis de chaque CSP en trois catégories d'âge 15-35 ans, 35-55 ans et plus de 55 ans ;
- sélectionner un homme et une femme de chaque tranche d'âge.

Une fois que notre corpus est construit, nous avons procédé à son enrichissement sur TXM[19] par l'ajout des couches d'annotation syntaxiques et sémantiques. Nous avons décrit la distribution du verbe au subjonctif, sa configuration rectionnelle ainsi que les traits morphosyntaxiques et sémantiques qu'il peut hériter de son gouverneur.

Résultats

Les différentes propriétés annotées ont constitué des variables pouvant faire l'objet de plusieurs requêtes possibles en rapport avec d'autres variables distributionnelles disponibles sur TXM (catégorie grammaticale, personne, etc.), ainsi qu'avec les métadonnées du corpus (ESLO1 *vs* ESLO2, genres interactionnels, tranche d'âge, niveau d'études, CSP, etc.).

Le versant quantitatif a fait apparaître en microdiachronie une baisse globale au niveau de l'usage du subjonctif en français parlé à Orléans. Cette perte de vitesse générale a été également soulignée par d'auteurs comme Grégoire & Bauche (1928), Frei (1929), Brunot & Bruneau (1947) et Riegel & al. (1994). Pour mieux comprendre cette donnée quantitative, nous nous sommes tournée vers les versants qualitatif et sociologique.

Les données qualitatives ont prouvé que cette chute ne touche réellement que l'emploi du subjonctif dans un contexte d'alternance modale et plus particulièrement dans les complétives affectées par les modalités interrogatives ou négatives (passage de 78 en ESLO1 % à 74 en ESLO2). Par contre, nous avons noté une légère augmentation du subjonctif quand il s'agit d'un cas d'emploi systématique excluant sa possible commutation avec un autre mode

[19] La description complète de ce logiciel est disponible sur ce lien https://txm.gitpages.huma-num.fr/textometrie/

(passage du 81,39 % en ESLO1 au 86,99 % en ESLO2), ce qui confirme ici sa *visibilité syntaxique* (Abouda, 2010)

Nous avons également relevé une quasi stabilité du subjonctif facultatif significatif[20] (passage du 3,88 % en ESLO1 au 2,14 % en ESLO2) ce qui confirme la rareté de ces cas d'emploi.

- ESLO2_ENT_1024_C, GC24, 0:35:09
  oui mais il est bien après qu'il **soit parti** de chez nous quoi avec des copains quoi

Ici, il est nécessaire de prendre en compte non seulement les compétences linguistiques mais aussi les critères sociologiques du locuteur. D'où l'intérêt de travailler sur le corpus ESLO.

L'omission du *que* rencontrée dans trois contextes d'emploi en ESLO1 a également attiré notre attention lors de l'analyse de nos données.

- ESLO1_ENT_001_C, OU, 0:59:30 :
  alors ça c'est dommage ça **vienne** à finir que je serais obligé de retourner

Bien qu'il s'agit d'un emploi marginal dans ESLO, il nous semble important d'alerter sur cette éventuelle évolution en cours qui est certes minime mais qui remet en question le rapport classique entre la conjonction *que* et le verbe au subjonctif.

Outre la structure morphémique, l'emploi des quatre paradigmes du subjonctif a fait couler beaucoup d'encre. Les résultats de notre corpus ont confirmé le recul du passé du subjonctif et la disparition de l'imparfait du subjonctif.

Au-delà des statistiques générales que nous venons d'exposer, nous avons examiné les traitements diaphasique et diastratique, susceptibles d'apporter plus d'éclaircissements sur l'usage des variables formelles généralement et du subjonctif en particulier, dans la mesure où les facteurs sociaux clarifient *ipso facto* la dynamique des pratiques langagières.

L'analyse de la variation situationnelle a permis de relever une baisse de 39 % dans l'emploi du subjonctif en entretiens semi-directifs, contexte plus ou moins formel ; une chute moins importante de 20 % dans les conférences universitaires, contexte strictement formel et une stagnation au niveau de l'emploi du subjonctif dans les repas, contexte informel. Ces résultats réfutent l'hypothèse selon laquelle le subjonctif est plus employé dans les cadres interactionnels formels et prouvent que l'impact de la variation situationnelle est véritablement peu significatif dans l'usage de cette forme verbale.

En examinant les proportions du subjonctif selon la catégorie d'âge, nous avons constaté en temps apparent une baisse significative de 50 % dans la catégorie de 15-35 ans ; une chute moins importante de 31% dans la catégorie de 55 ans et une stabilité dans la catégorie de 35-55 ans. D'après ces chiffres, nous concluons que le changement en cours produit au niveau de l'usage du subjonctif, n'est pas forcément rattaché à la stratification en âge.

Nous avons aussi dégagé une diminution au niveau de l'emploi du subjonctif par la catégorie la plus élevée sur l'échelle sociale, celle des cadres ; une baisse moins signifiante chez la catégorie sociale moyenne, celles des employés et une certaine stabilité dans l'emploi du subjonctif par la catégorie sociale la plus faible, celle des ouvriers. Nous avons également

[20] Il s'agit d'un emploi non obligatoire du subjonctif. Aucun élément grammatical n'exige l'utilisation de ce mode verbal dans ce genre de contexte.

remarqué une grande variation dans l'emploi du subjonctif par toutes les tranches d'âge quelles que soient leurs catégories socioprofessionnelles. Ces changements attestent que l'emploi du subjonctif ne se rapporte pas aux facteurs sociaux.

Conclusion

Par le biais de cette étude, nous avons pu observer le comportement du subjonctif tel qu'il s'actualise dans les conversations orales. L'empan diachronique saisi par notre corpus, combiné au caractère situé des données, nous a offert la possibilité d'étudier les différents indices de la variation dans son usage en français parlé à Orléans. Nos résultats ont manifesté que les variables externes parlent peu, face à la systématicité de ce mode.

Références bibliographiques

Abouda, L. (2010). De la visibilité syntaxique des modes, de l'invisibilité syntaxique des temps. *Liens linguistiques. Etudes sur la combinatoire et la hiérarchie des composants*, 319-333.

Abouda, L., & Skrovec, M. (2018). Pour une micro-diachronie de l'oral : le corpus ESLO-MD. In *SHS Web of Conferences* (Vol. 46, p. 11004). EDP Sciences.

Ferdinand, B., & Bruneau, C. (1937). *Précis de grammaire historique de la langue française*. Paris : Masson.

Frei, H. (1929). La grammaire des fautes. Introduction à la linguistique fonctionnelle. *Paris-Genève, Geuthner*.

Grégoire, A. & Bauche, H. (1928). Le langage populaire. *Revue belge de Philologie et d'Histoire*. Vol. 7, no. 4, p. 1536-1541.

Heiden, S., Magué, J. P., & Pincemin, B. (2010, June). TXM : Une plateforme logicielle open-source pour la textométrie-conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010* (Vol. 2, No. 3, pp. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto.

Kastronic, L. (2016). *A comparative variationist approach to morphosyntactic variation in Hexagonal and Quebec French* (Doctoral dissertation, Université d'Ottawa/University of Ottawa).

Laurier, M. (1989). Le subjonctif dans le parler franco-ontarien : un mode en voie de disparition. *Le français canadien parlé hors Québec : aperçu sociolinguistique*, 105-126.

Nølke, H. (1985). Le subjonctif : fragments d'une théorie énonciative. *Langages*, (80), 55-70.

Nordahl, H. (1969). *Les systèmes du subjonctif corrélatif : étude sur l'emploi des modes dans la subordonnée complétive en français moderne* (No. 1). Universitetsforlaget.

Riegel, M., Pellat, J. C., & Rioul, R. (1994). Grammaire méthodique du français. *Linguistique nouvelle*.

Soutet, O. (2000). *Le subjonctif en français*. Editions Ophrys.

Tesnière, L. (1959). Eléments de syntaxe structurale. *Paris, Klincksieck*, 25.

L'apport des ESLO pour la documentation du continuum linguistique dans le petit Maghreb

Ben Barka Messaoudi Fatma [1], Rayan Ziane [2] et Anissa Aissani [2]
[1] Laboratoire EMA, CY Cergy Paris Université INSPE [2] Laboratoire CRISCO, Université Caen
fatma.messaoudi1@cyu.fr, rayan.ziane@unicaen.fr, anissa-fella.aissani@etu.unicaen.fr

Introduction

Au cours de ces dernières années, les avancées informatiques ne cessent de renouveler les recherches linguistiques. Comme le note Baude (2007 : 85), les toutes nouvelles technologies de stockage, de diffusion mais aussi d'exploitation des enregistrements sonores, couplées aux outils (transcriptions synchronisées sur le signal, annotation, etc.) ouvrent des perspectives prometteuses pour les études sur les corpus de langues parlées. Dans ce mouvement collectif de collecte et de diffusion de données orales, le Laboratoire Ligérien de Linguistique a joué un rôle important par la mise à disposition des enquêtes sociolinguistiques à Orléans (ESLO) à tout.e chercheur/euse intéressé.e. Ce corpus dispose de trois principaux atouts : 1) des données en masse (10 000 000 mots), 2) des données situées (métadonnées informant sur le profil du locuteur en termes d'âge, de sexe et de la catégorie socioprofessionnelle) et 3) des données micro-diachroniques (ESLO1 1968, ESLO2 à partir de 2008). Cette sélection de données tenant compte de plusieurs variétés langagières nous a motivés à construire un grand corpus comparable en arabe maghrébin.

Situation linguistique dans le petit Maghreb

Le paysage linguistique dans le petit maghreb se caractérise par la coexistence de l'arabe standard moderne (ASM) et l'arabe parlé. Bien que les parlers tunisien, algérien et marocain constituent les langues maternelles des locuteurs maghrébins, ils ont été largement déconsidérés et mis à l'écart. En effet, l'idéologie arabo-musulmane a longtemps occasionné la non reconnaissance de toute forme d'évolution linguistique en considérant que ces parlers sont le résultat de la décadence dues au contact. À l'exception du contexte médiatique et plus particulièrement du domaine publicitaire, ces variétés, privées de descriptions grammaticales et d'orthographes stables, sont exclues des programmes d'éducation, des écrits administratifs et parfois des émissions télévisées au profit de l'ASM.

Ne bénéficiant ni de statuts officiels, ni de ressources linguistique[21], ces langues sous-documentées manquent de données et d'outils en libre accès facilitant leurs descriptions sans tomber dans les pièges de l'intuition. Ayant la volonté de faire avancer le débat sur les enjeux théoriques et pratiques de la documentation et de la description des langues peu dotées,

[21] Hormis quelques tentatives récentes de recueil des parlers tunisien et marocain (Graja *et al.* 2010, Zribi *et al* 2015, Moukrim 2010 ...) s'inscrivant dans le sillage du développement général de la linguistique de corpus et du traitement automatique du langage.

nous avons exploité la démarche de la constitution d'ESLO pour constituer notre grand corpus d'arabe maghrébin[22].

Méthodologie

Exploitation de la méthodologie d'ESLO

Le réservoir ESLO illustre un ensemble de bonnes pratiques à préconiser pour minimiser la complexité des procédures de collecte et de transformation des données primaires en données secondaires. Dans la mesure du possible, nous les avons suivies afin de construire un corpus partageable et interopérable. Nous avons favorisé comme mode de collecte des données l'entretien semi-dirigé en face-à-face, "situation certes très formelle, mais qui avait l'avantage d'être (…) contrôlable." (Abouda et Baude, 2006 : 4). Notre corpus compte aussi un certain nombre d'entretiens réalisés via des plateformes de visio-conférences (en particulier pour la partie algérienne du corpus). Le guide d'entretien a été puisé dans les principaux thèmes retenus par ESLO (logement, travail, loisirs, langues, ville d'habitation), tout en ajoutant d'autres thématiques (sur la révolution tunisienne et sur les représentations langagières en Algérie) susceptibles de faire parler les locuteurs tunisiens et algériens.

Afin de maintenir un certain équilibre au sein de notre corpus, deux autres genres interactionnels de « contrôle », i.e. les repas et les conférences universitaires sont et seront intégrés, à hauteur de 20%, tout en respectant les exigences de nos terrains. En ce sens, quelques aménagements ont dû être faits lors du recueil des conférences universitaires. En effet, nous nous sommes rendus compte que, dans ce contexte formel, l'arabe parlé cède sa place à l'ASM. Nous avons donc décidé de remplacer les conférences des ESLO par des cours universitaires. Néanmoins, cette décision n'a pas été suffisante pour résoudre ce problème étant donné que la plupart des cours universitaires privilégient la variété standard. Nous nous sommes donc retournés vers les cours artistiques (de musique, de danse et de dessin) où les professeurs ont plus de liberté pour s'exprimer dans la langue vernaculaire du pays.

D'autres choix méthodologiques et techniques ont été opérés afin de faire face aux contraintes rencontrées sur terrain. Nous pouvons citer à titre d'exemple, le changement du lieu de l'enquête (d'Orléans en Tunisie) ou le recours à des enregistrements en visio-conférence (données algériennes) pour certains enregistrements afin de gérer le manque de quelques profils sociologiques ou le déséquilibre intérieur en termes de variations diaphasique et diatopique.

Dans l'objectif de représenter la quasi totalité du continuum linguistique maghrébin, nous avons également opté pour un échantillonnage des locuteurs en nous basant sur les principes suivants :

- nous reposer sur trois catégories socioprofessionnelles (CSP) : cadres, employés, ouvriers ;

[22] Deux premières enquêtes en parlers tunisien et algérien ont été déjà entamées dans le cadre d'une étude doctorale intitulée Étude contrastive du subjonctif en français parlé à Orléans et de ses éventuels équivalents en arabe tunisien (2022) et d'un mémoire de recherche (en cours) portant sur les représentations langagières en Algérie. L'objectif suivant est de mener la troisième enquête en arabe marocain.

- classer les locuteurs sélectionnés de chaque CSP en trois tranches d'âge 15-35 ans, 35- 55 ans et plus de 55 ans ;
- choisir un homme et une femme de chaque catégorie d'âge.

Données situées protégées

Conscients des avantages offerts par les données situées dans le traitement des faits linguistiques, nous avons veillé à récolter tous les renseignements concernant nos enregistrements ainsi que leurs contextes de production en mettant en place deux formulaires :

- un formulaire Témoin comportant des informations sur l'âge, le sexe, le niveau scolaire, la pro- fession, la CSP, les langues parlées et lieu de naissance.
- un formulaire Enregistrement informant sur la situation de communication enregistrée, sa date, son lieu, sa durée et le nombre de participants.

Afin de protéger nos locuteurs et nos données des éventuels problèmes de collecte, de transmission et de la propriété intellectuelle, nous avons demandé aux témoins la signature d'un texte de consentement écrit synthétisant le cadre et les finalités de nos enquêtes.

Traitement

L'absence de standardisation du code orthographique pour les parlers maghrébins nous a poussés à questionner les pratiques opérées de façon systématique dans les transcriptions du corpus ESLO. Là où une transcription phonétique serait trop coûteuse en termes de temps, nous avons adopté une transcription morphophonologique, proche des formes usuelles répandues sur les réseaux sociaux communément caractérisées par le terme Arabizi (Yaghan, 2008). C'est aussi par souci d'universalisation des données que nous avons eu recours à la graphie latine et aux caractères relatifs à l'API permet- tant de combler les lacunes du premier alphabet, en suivant les conventions de l'Institut National des Langues et Civilisations Orientales (INALCO). Néanmoins, pour codifier les spécificités de l'oral, nous nous sommes servis des propositions d'ESLO. Tenant compte des faveurs des outils conçus pour le traitement automatique de la langue arabe et des méthodes d'apprentissage automatique, nous avons ensuite réalisé une translittération automatique des données transcrites en caractères latins vers les caractères arabes, grâce à l'outil API de Google Input et le translittérateur ATAR (Talafha et al. 2021).

De la diffusion à l'archivage

Notre projet s'inscrit dans une démarche de science ouverte et de diffusion des données de la recherche, ainsi nous prévoyons de mettre à disposition les transcriptions et toutes formes d'annotations ajoutées. Toutefois, les enregistrements audio constituent des données à caractères personnels sensibles, ce qui exclut leur diffusion publique. Par ailleurs, un archivage des données pérenne est planifié à l'issue de la réalisation du projet afin de conserver ce paysage sonore des parlers du petit Maghreb.

Conclusion

Face aux innombrables "corpus fantômes" cités dans la littérature sur les parlers du petit Maghreb, nous proposons le premier corpus "équitable" pour ce continuum linguistique. Équitable étant donné qu'il répond aux principes FAIR (D. Wilkinson et al, 2016) qui doivent

être opérés comme une ligne directrice pour tout projet qui vise à partager et à rendre accessible les données de la recherche. Notre corpus sera aussi Findable/facile à trouver, cela implique qu'il se doit d'être stocké dans un site/plateforme scientifique et décrit par des métadonnées pour faciliter les recherches. Notre corpus sera également Accessible et ouvert, en accès libre, à tous les acteurs et actrices de la communauté scientifique. Findable et Accessible via la plateforme HUMA-NUM, nos données seront aussi Interopérables car elles seront sous format largement diffusé et faciles à lire et à traiter, au format XML dans notre cas. Notre corpus sera ensuite Reusable/réutilisable puisque nous permettrons son partage et sa réutilisation.

Enfin, ce corpus a pour vocation à être doté d'une annotation morphosyntaxique sous la forme de treebank en nous inspirant de la méthodologie proposée par Kahane *et al.* (2021) dans les formalismes Universal Dependencies (De Marneffe *et al.*, 2021) et Surface Syntactic Universal Dependencies (Gerdes *et al*, 2018).

Références bibliographiques

Abouda, L., Baude, O. (2006). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO. *In Corpus en Lettres et Sciences sociales, Des documents numériques à l'interprétation.*

Baude, O. (2007). Aspects juridiques et éthiques de la conservation et de la diffusion des corpus oraux. *Revue française de linguistique appliquée*, (1), 85-97.

Ben Barka Messaoudi, F. *Étude contrastive du subjonctif en français parlé à Orléans et de ses éventuels équivalents en arabe tunisien.* PhD Thesis. Université d'Orléans. 2022.

Darwish, K., Attia, M., Mubarak, H., Samih, Y., Abdelali, A. (2020). Effective Multi Dialectal Arabic POS Tagging. *Natural Language Engineering*, 1(1), 18.

De Marneffe, M. C., Manning, C. D., Nivre, J., Zeman, D. (2021). Universal dependencies. *Computational linguistics, 47* (2), 255-308.

Gerdes, K., Guillaume, B., Kahane, S., Perrier, G. (2018, novembre). SUD or Surface-Syntactic Universal Dependen- cies : An annotation scheme near-isomorphic to UD. *Universal Dependencies Workshop 2018*.

Kahane, S., Vanhove, M., Ziane, R. Guillaume, B. (2021). A morph-based and a word-based treebank for Beja. *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria.

Talafha, B., Abuammar, A., Al-Ayyoub, M. (2021). ATAR : Attention-based LSTM for Arabizi transliteration. *International Journal of Electrical and Computer Engineering, 11* (3), 2327.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3* (1), 1-9.

Yaghan, M. A. (2008). " Arabizi" : A Contemporary Style of Arabic Slang. *Design Issues, 24* (2), 39-52

Recueillir et utiliser des corpus en crèche : une recherche collaborative avec les professionnelles de la petite enfance

Tiphanie Bertin[1], Caroline Masson[1], Christine da Silva Genest[2], Roxane Perrin Hennebelle[1]
[1]EA 7345 CLESTHIA, Université Sorbonne Nouvelle
[2]EA 3450, DevAH, Université de Lorraine
tiphanie.bertin@sorbonne-nouvelle.fr, caroline.masson@sorbonne-nouvelle.fr,
christine.da-silva-genest@univ-lorraine.fr, roxane.perrin-hennebelle@sorbonne-nouvelle.fr

Introduction

Durant leurs trois premières années, les enfants expérimentent différents contextes communicationnels, dépendants notamment de leur mode de garde : à domicile (et/ou au sein de la famille élargie), chez une assistante maternelle ou dans un établissement d'accueil du jeune enfant (EAJE). Ces structures d'accueil collectif sont perçues comme des espaces privilégiés pour l'acquisition du langage, offrant des situations d'interaction variées (dyadiques, polyadiques, entre adultes et enfants, entre enfants…). Des bénéfices pour le développement du langage de l'enfant sont rapportés, en particulier dans les pays où la qualité de l'accueil collectif est au centre des politiques éducatives (Sales-Cordeiro et al., 2020). Pour le contexte français, Marcos et collègues (2004) ont notamment mis en évidence des compétences structurelles et fonctionnelles plus avancées chez les enfants gardés en crèche par rapport à ceux gardés par leur mère ou chez une assistante maternelle (longueur moyenne d'énoncés plus élevée, tours de parole et énoncés par tour de parole plus nombreux).

Les professionnelles de la petite enfance sont sensibles au langage des enfants et à celui qu'elles leur adressent. Ainsi, elles mettent en œuvre des pratiques pouvant soutenir les enfants dans leur apprentissage du langage (Degotardi, 2021 ; Girolametto et al., 2000). Or, ces conduites sont menées de façon plus ou moins consciente par les professionnelles (Masson & Bertin, 2021). En outre, bien que leurs formations initiales et continues mettent en avant leur rôle pour accompagner le développement langagier des jeunes enfants, celles-ci ne leur donnent pas les clés pour comprendre les effets de leurs pratiques langagières sur l'acquisition des aspects structurels, fonctionnels et communicatifs du langage de l'enfant (Masson & Bertin, à paraître). C'est ainsi que ces professionnelles expriment des besoins de connaissances et de développement professionnel sur l'acquisition du langage pour compléter leurs formations initiales ou valider leurs intuitions (Salvi & Prince Agbodjan, 2021). Les linguistes spécialistes de l'acquisition sont en mesure de répondre à leurs demandes en raison de leur expertise dans le domaine.

Mais la réussite de cette collaboration entre linguistes et professionnelles est dépendante des modalités mises en œuvre. Elle ne peut fonctionner que dans un partage d'expertise, la formation descendante étant moins efficiente dans ce genre de cas. L'un des moyens mobilisés est le recours aux traces de l'agir professionnel par l'utilisation de corpus de données attestées, recueillies sur le terrain des personnes concernées. Les corpus permettent ainsi une

meilleure connaissance du terrain pour les linguistes et un matériau de réflexion concret pour les professionnelles.

Alors que les corpus d'interactions parents-enfants sont nombreux et permettent de décrire l'expérience communicative en famille, les caractéristiques de l'environnement langagier en structures d'accueil collectif restent peu documentées (Marcos et al., 2004). Pour les chercheur·ses qui explorent ce terrain, le défi est d'appréhender la diversité des interactions dans ces établissements. Une inscription dans le cadre de la "linguistique impliquée" (Canut, Husianycia & Masson, 2021), avec une posture active et médiatrice et pas seulement observatrice, permet d'analyser les pratiques langagières des enfants et des adultes durant les activités quotidiennes, tout en profitant de l'expertise des professionnelles.

Le travail que nous présentons ici vise à produire et transmettre des connaissances sur le langage des enfants et des adultes en EAJE, ses usages et ses modalités d'apprentissage tout en développant des actions permettant à la recherche et à la pratique de collaborer. Les travaux des linguistes sur le développement langagier et les interactions adultes-enfants (Bruner, 1983 ; Tomasello, 2003) constituent la base du travail conjoint avec les professionnelles qui s'appuient sur ces connaissances pour développer une expertise, élaborée à partir de leurs propres savoirs et savoir-faire.

Notre communication présentera les modalités de cette collaboration et plus particulièrement celles de l'analyse conjointe des données entre les chercheur·ses et les professionnelles sur les conduites langagières en fonction du type d'activité (Hoff-Ginsberg, 1991 ; Salazar Orvig et al., 2018). En effet, les pratiques langagières, tant leur forme que leur fonction, sont fortement influencées par les activités en cours. Celles-ci se caractérisent donc par des cadres privilégiés dans lesquels l'enfant apprend à dialoguer et à produire des formes langagières adaptées. Nous montrerons comment les outils d'analyse de la linguistique peuvent être mobilisés et adaptés pour ce public de non-spécialistes afin de les aider à décrire et à comprendre leurs actions en situation.

Corpus et méthodologie

Corpus

Notre réflexion s'appuie sur des données recueillies dans le cadre d'un projet de recherche mené en EAJE. Plus de 260 enfants (de 4 mois à 3 ans ½) et 80 professionnelles (auxiliaires de puériculture et éducatrices jeunes enfants) sont impliqué·es dans le projet. Le corpus est constitué de 79h de vidéos d'interactions entre des professionnelles et des enfants et des enfants entre eux dans des activités variées et caractéristiques de ces structures (ex. repas, jeux libres, jeux pédagogiques, lectures, motricité). Une partie de ces enregistrements (environ 12h) est transcrite sous CLAN (MacWhinney, 2000).

Méthodologie

A partir d'une sélection d'enregistrements recueillis dans le cadre du projet, nous présenterons notre approche pour accompagner les professionnelles dans leurs observations et réflexions sur le langage des enfants et leurs propres conduites langagières. Nous suivons la démarche de la linguistique de corpus, du recueil à la description des faits linguistiques (Lejeune, 2010), considérant que travailler sur des données attestées et analysées constitue un point de départ de la prise de conscience des professionnelles sur leurs pratiques langagières (da Silva Genest

& Masson, 2017). Nous discuterons ainsi des choix des chercheur·ses lors de la constitution d'un corpus à double objectif : celui de décrire et documenter les interactions en EAJE et celui de mener un transfert d'expertise vers des professionnelles de la petite enfance, ce qui nécessite de rendre les résultats de la recherche exploitables pour le terrain. Nos analyses portent sur la description des aspects structurels (ex. longueur moyenne des énoncés, diversité lexicale, complexité syntaxique) et fonctionnels (ex. occupation de l'espace discursif, mouvements illocutoires, inscription dans les échanges) des productions des enfants et des adultes, ainsi que sur les conduites d'étayage (ex. reprises et reformulations, questions) et leurs effets sur les enfants compte tenu du type d'activité (jeux, repas, lectures). Après une présentation de nos résultats, nous exposerons les modalités de l'accompagnement des professionnelles dans le réinvestissement de ces analyses qui servent de base au travail collaboratif.

Résultats

La démarche repose sur deux étapes : dans un premier temps, une analyse linguistique des données recueillies dans les structures est réalisée, puis dans un second temps, ces données et résultats sont explorés avec les professionnelles.

Nous présenterons tout d'abord les résultats obtenus pour l'analyse des aspects linguistiques structurels et fonctionnels des activités sélectionnées : un jeu pédagogique, un repas et une narration. Ces mesures nous permettront de mettre en évidence les conduites des professionnelles et des enfants compte tenu de l'activité. Par exemple, les analyses menées sur l'activité de narration montrent que les formes et les usages mobilisés permettent aux enfants de développer des savoirs et des savoir-faire propres à la narration comme la dénomination, la description ou encore la production de formes langagières complexes. Dans les activités de jeu et de repas d'autres usages langagiers (ex. ordres, projections) sont mobilisés et la productivité langagière est plus réduite. Les activités sont ainsi tournées vers des objectifs qui ne visent pas toujours en premier lieu la production ou la compréhension du langage (comme dans le cas de la narration) mais, par exemple, l'apprentissage des règles en collectivité (partage, contrôle de la frustration…), comme dans les activités de jeu et de repas. En cela, elles impactent les pratiques langagières que les enfants expérimentent.

Nous exposerons ensuite les réflexions et les résultats du travail collaboratif mené avec les professionnelles. Nous discuterons des modalités de transmission des analyses linguistiques et notamment de la mise en place de retours sur pratiques réguliers. Nous aborderons la question du support à utiliser lors de ces retours, des apports et limites de l'utilisation d'extraits vidéos ou de transcriptions. La description du travail conjoint d'analyse de séquences du corpus permettra de décrire la démarche du transfert d'expertise progressif auprès des professionnelles, avec une présentation de l'évolution de leur démarche analytique à mesure des retours sur pratiques. Ces dernières sont par exemple beaucoup plus sensibles aux questions d'occupation de l'espace discursif et d'adressage qu'à d'autres résultats dont les chercheur·ses ont besoin (ex. nombre de mots par minute, types d'énoncés).

Dans cette perspective, nous décrirons notre démarche pour répondre aux demandes d'outils d'observation (nouveaux ou complémentaires) exprimées par les équipes. Nous présenterons ainsi deux exemples d'outils élaborés dans le cadre du projet. Le premier, une grille d'observation des manifestations communicatives et langagières des bébés pour un suivi longitudinal des enfants, a vocation à répondre à un besoin exprimé par les professionnelles de noter/de consigner des aspects saillants du développement langagier des enfants, et

transmettre ces observations lors du changement de sections. Le second outil, un inventaire des productions récurrentes des professionnelles, émane à l'inverse d'un besoin des chercheur·ses pour mieux appréhender l'input reçu par les enfants, et notamment les routines langagières.

En parallèle, nous interrogerons l'utilisation des corpus pour travailler avec des professionnelles qui ont des connaissances diverses sur le langage de l'enfant mais qui n'ont pas été formées à la linguistique dans leurs parcours professionnels.

Pour terminer, cette communication nous donnera l'occasion de présenter les apports de ce corpus inédit dans un cadre éducatif peu décrit et ses perspectives de travail pour la recherche en acquisition du langage.

Références bibliographiques

Bruner, J. S. (1983). *Le développement de l'enfant: Savoir-faire, savoir dire*. Paris (PUF).

Canut, E., Husianycia, M., & Masson, C. (2021). Une linguistique impliquée en acquisition du langage ?. *Études de linguistique appliquée*, *202* (2), 141-153.

da Silva Genest, C. & Masson, C. (2017). Apport de la linguistique de corpus à l'étude des situations cliniques: utilisation de ressources écologiques pour évaluer les pratiques professionnelles. *Studii de Lingvistică*, 7, 89-112.

Degotardi, S. (2021). The language environment of infant child care. Issues of quantity, quality, participation and context. In O. Saracho, *Contemporary Perspectives in Early Childhood Education*. Charlotte, NC: IAP, 85-107.

Girolametto, L., Hoaken, L., *et al.* (2000). Patterns of adult-child linguistic interaction in integrated day care groups. *Language, Speech, and Hearing Services in Schools*, *31(2),* 155-168.

Hoff-Ginsberg, E. (1991). Mother-child conversation in different social classes and communicative settings. *Child Development*, *62*, 782-796.

Lejeune, C. (2010). Montrer, calculer, explorer, analyser. Ce que l'informatique fait (faire) à l'analyse qualitative, *Recherches Qualitatives* 9, 15-32.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ (Lawrence Erlbaum Associates).

Marcos, H., Salazar Orvig, A., et *al.* (2004). *Apprendre à parler : influence du mode de garde*. Paris : L'Harmattan.

Masson, C. & Bertin, T. (2021). Accompagner le développement langagier du très jeune enfant : Une recherche-action-formation dans les structures de petite enfance. *TRANEL, 74*, 31-45.

Masson, C. & Bertin, T. (à paraître). De la recherche à la pratique et vice-versa : pour une linguistique impliquée avec des professionnel·les de la petite enfance. In M.A. Akinci, I. Maillochon & V. Miguel-Addisu (Eds), *Vivre et parler avec le jeune enfant en crèches multi-accueil*. Paris : L'Harmattan.

Salazar Orvig, A., Marcos, H., *et al.* (2018). Referential features, speech genres and activity types. In M. Hickmann, H. Jisa & E. Veneziano (éds.), *Sources of variation in first language acquisition: Languages, contexts, and learners.* Amsterdam (Benjamins), 219-242.

Sales Cordeiro, G., Zogmal, M., et *al*. (2020). *Projet CapLang: soutenir les capacités langagières et littéraciques des enfants à travers le développement professionnel des éducatrices et éducateurs de l'enfance.* Forum-lecture.ch, Plate-forme en ligne pour la littératie, 1/2020.

Salvi, V. & Prince Agbodjan, B. (2021). Le langage, les langues et la culture : des objets d'attention conjointe en crèche. *Spirale*, 99-3, 152-157.

Tomasello, M. (2003). *Constructing a language. A usage-based theory of language acquisition.* Cambridge (Harvard University Press).

Caractériser le discours de l'accès aux droits : quels corpus pour quels résultats ?

Marie Bouchet
Laboratoire CLILLAC-ARP, Université Paris Cité
marie.bouchet@u-paris.fr

Introduction

Depuis le début des années 2000, la société vit une transition numérique globale. Cette transition est également présente dans l'administration, et notamment dans les relations entre les citoyens et leurs gouvernements. La dématérialisation des démarches administratives marque une modification radicale des rituels d'interactions entre les administrations et les administrés, ce qui impacte naturellement les discours. Le discours de l'accès aux droits, qui constitue l'objet principal de notre étude, est notamment bouleversé, car il est recentré sur un seul média : le site web. Désormais, les informations officielles sont communiquées par les sites internet et sur les espaces utilisateurs en ligne, via email par exemple (Lochak, 2022). Le discours de l'accès aux droits est central dans une société, puisqu'il concerne une grande majorité des citoyens. Ses caractéristiques lui confèrent sa spécificité tout en dressant certains obstacles à la communication, et donc à l'accès aux droits : il s'agit d'un discours spécialisé régulièrement simplifié ou vulgarisé, qui demeure néanmoins complexe. Il est aujourd'hui presque toujours numérique, qu'il soit imprimé, disponible sous un format statique ou dynamique, en ligne ou local. La transition numérique, par son impact sur la langue et le discours, amène de plus en plus de linguistes (Paveau 2015 ; Simon, 2018 ; Souchier et al., 2019) à aborder le discours numérique en tant que forme ou variante spécifique de la langue. Nous souhaitons dans notre étude caractériser ce nouveau discours administratif afin de prendre en considération ses enjeux discursifs dans la société, dans deux pays : le Royaume-Uni et la France. Ces pays ont connu une évolution similaire dans deux langues différentes, l'anglais et le français, dont les discours ont évolué de manière spécifique. L'objectif de cette étude est de donner des clés pour aider à la compréhension de ce discours, notamment sur le plan linguistique et discursif (en lien avec la langue de spécialité de l'administration, sa terminologie et sa phraséologie), tout en y intégrant sa dimension numérique qui est, à notre avis, centrale. Cette étude sera enrichie par une étude du contexte social et culturel de ce discours, en prenant en compte son histoire, mais aussi les recherches réalisées dans d'autres domaines (sociologie, rédactologie, sciences politiques, etc.). L'objectif de cette caractérisation est également de participer aux efforts de visibilité de ce discours dans la recherche, afin que la communauté scientifique et administrative puisse se saisir pleinement de ces enjeux linguistiques et sociaux.

Pour caractériser ce discours spécialisé, nous avons étudié les textes publiés sur les sites administratifs officiels, principalement service-public.fr pour la France, et gov.uk pour le Royaume-Uni, en construisant des corpus à partir de sites web. Afin de prendre en compte les particularités de notre discours et les objectifs de notre étude, les choix méthodologiques, notamment en termes de constitution du corpus, se sont avérés essentiels. Pour caractériser ce type de discours, une méthodologie fondée sur l'élaboration et l'analyse de deux types de corpus a été mise en place et explorée. Dans la première partie de cette communication, nous

présenterons les particularités d'un discours spécialisé issu de sites web. Dans la deuxième partie, nous montrerons comment nos observations ont mené à l'élaboration de deux types de corpus complémentaires. Nous conclurons avec les résultats d'analyses issus de l'étude combinée de ces deux types de corpus et les interrogations que ces résultats soulèvent.

Le discours numérique de l'accès aux droits

Le discours étudié est nativement numérique : il s'agit de « productions élaborées en ligne, dans les espaces d'écriture et avec les outils proposés par internet » (Paveau, 2017 : 27). Notre discours en particulier est rédigé sur un site web dynamique, ce qui signifie qu'il contient des pages générées à la demande par le serveur, selon les requêtes de l'utilisateur. L'utilisateur du site web a donc une liberté d'interaction avec le site, qui s'adapte et affiche les informations demandées. C'est le cas pour le site service-public.fr, qui est composé de nombreux éléments interactifs, comme on peut le voir sur l'image ci-dessous.

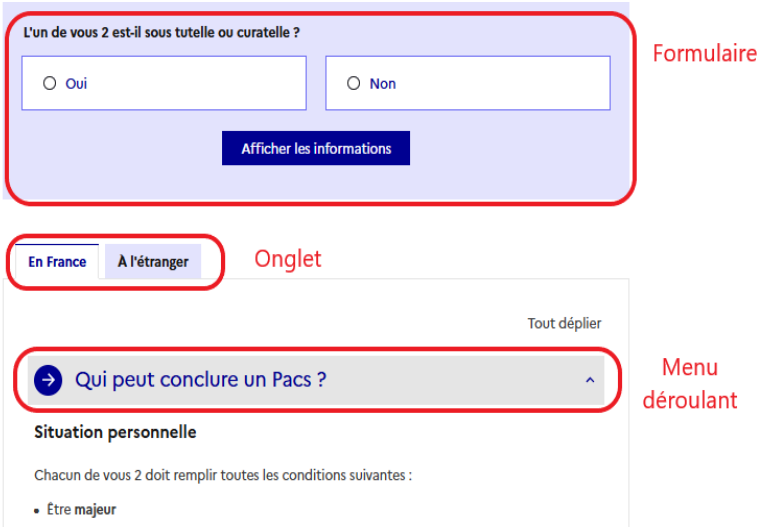


figure . 1 Page web avec éléments interactifs

La figure 1 provient de la page « Se pacser » du site officiel de l'administration française (service-public.fr). Elle illustre certains types d'éléments interactifs présents sur les pages du site : les formulaires, les onglets et les menus déroulants. En plus de ces éléments interactifs, on retrouve également divers menus qui permettent de naviguer sur le site web. Ces discours sont délinéarisés : le fil du discours est fragmenté par les éléments techniques (Paveau, 2017) et les différents modules présents sur la page (Maingueneau, 2016). Le fil du discours devient un objet navigable, et explorable, afin de trouver l'information recherchée. Le domaine de l'accès aux droits, s'actualisant désormais essentiellement dans le discours numérique, est également vecteur de connaissances spécialisées et relève donc tout autant du discours spécialisé (Lerat, 1995). Il est le résultat d'une activité de diffusion vers l'extérieur de connaissances détenues par une communauté experte, il s'agit donc d'une activité de vulgarisation (Authier, 1982).

Notre étude vise également à envisager le site web de l'accès au droit gouvernemental en tant que genre. D'après la définition de Swales (1990 : 58), un genre est « a class of communicative events, the members of which share a certain set of communicative purposes ». Ces sites correspondent en effet à des objectifs qui sont la diffusion de l'information administrative et juridique aux usagers sur un support numérique et interactif.

Plusieurs chercheurs remarquent également qu'un genre est défini par une certaine stabilité (Bakhtine, 1984 ; Adam, 1997 ; Delavigne, 2022), mais que ce sont également des « communications conventionnées ». Les évolutions des dernières décennies ont mis en avant les plateformes numériques pour la communication avec l'État. Ces communications répondent maintenant à un certain nombre de normes, mais aussi de régularités linguistiques. Le site web de l'accès aux droits s'est ainsi positionné en tant que genre destiné à la communication entre l'administration et les administrés concernant leurs droits.

Méthode d'élaboration des corpus

Les critères de constitution d'un corpus doivent être définis en accord avec les objectifs de recherche. Les corpus servant à caractériser le discours de l'accès aux droits doivent prendre en compte ses particularités : discours spécialisé, numérique et vulgarisé. Pour cela, nous avons décidé d'élaborer deux ensembles de corpus complémentaires. Le premier ensemble répond aux critères d'élaboration des corpus dans la tradition de Sinclair (1991). Les corpus sont de grande taille et ils sont compilés et annotés automatiquement en catégorie syntaxique. La table 1 propose un récapitulatif de ce corpus qui est bilingue : français et anglais britannique. Il est divisé en 3 thématiques que l'on retrouve sur les sites d'accès aux droits.

	Français (FR)	Anglais (R-U)
Aides sociales	722 000 mots — 4 sites	211 000 mots — 6 sites
Citoyenneté	711 000 mots — 4 sites	303 000 mots — 4 sites
Immigration	1 233 000 mots — 3 sites	124 500 mots — 1 site

table 1. : Récapitulatif de l'ensemble de corpus 1

Ce corpus permet de faire des explorations notamment phraséologiques et terminologiques du discours. L'annotation en catégorie syntaxique et l'exploration dans des logiciels permettent l'étude de termes et des structures collocationnelles ou routines discursives. Dans ce tableau, nous pouvons observer une différence de taille considérable entre les corpus. Le corpus français comporte beaucoup plus de mots que le corpus anglais. Cela est dû à la politique de communication des deux pays. Au Royaume-Uni, la communication est centralisée autour d'un site : gov.uk et d'une organisation : *Government Digital Service*. En France, la Direction de l'information légale et administrative (DILA) est l'organisme principal de la diffusion d'informations administratives, notamment sur le site service-public.fr. Il existe cependant d'autres entités (Caisse d'allocations familiales, Pôle emploi, etc.) externes au gouvernement responsable d'une certaine partie de la communication. Il existe également d'autres initiatives de diffusion des informations concernant l'accès aux droits, comme le site web Welcome to France, un service d'information bilingue (français et anglais) à destination des étrangers et de leur famille. De plus, les démarches sont moins nombreuses au Royaume-Uni qui a construit une structure plus centralisée que la France, qui compte différents organismes suivant les types de services : aides sociales, droit au séjour, logement, etc. L'emploi généralisé du *plain language* sur le site britannique peut également expliquer ces disparités.

Nous avons également créé un corpus de petite taille à partir du parcours utilisateur. Ces petits corpus sont créés à partir d'un nombre de pages qui correspondent à un parcours fictif de navigation. La table 2 propose un récapitulatif de ce corpus qui est bilingue : français et anglais britannique. Il est composé de trois parcours utilisateurs par langue correspondant aux démarches suivantes : l'aide au logement, l'acte de naissance et la naturalisation.

	Français (FR)	Anglais (R-U)
Aide au logement	13 314 mots — 5 pages	4 500 mots — 5 pages
Acte de naissance	25 937 mots — 8 pages	5 400 mots — 5 pages
Naturalisation	29 564 mots — 6 pages	7 368 mots — 8 pages

table 2. : Récapitulatif de l'ensemble de corpus 2 (parcours)

Chaque page de parcours est annotée manuellement avec les fonctions des différents segments, en s'inspirant de la théorie des moves, qui sont définis comme des « discourse segments that can move discourse forward by performing coherent communicative functions » (Swales, 2004). Enfin, le corpus prend en compte la dimension dynamique et interactive du discours en annotant manuellement les hyperliens. Dans ce second ensemble de corpus, nous pouvons également remarquer une disparité entre le nombre de mots dans les deux langues. Le corpus français est plus conséquent, tandis que le nombre de pages récoltées est similaire dans les deux langues. Cette différence vient de la multiplicité des démarches et étapes à réaliser dans un contexte français, tandis qu'il n'existe généralement qu'une démarche, avec peu d'étapes dans le contexte britannique. Pour l'aide au logement par exemple, il en existe trois types en France, qui correspondent à différentes conditions. L'aide au logement doit également être demandée sur le site de la Caf, ce qui sous-entend de rediriger l'utilisateur vers un site externe au site gouvernemental. Au Royaume-Uni, l'aide au logement fait partie de l'aide intitulée *Universal Credit*, un dispositif unique qui rassemble une grande partie des allocations (logement, familiales, recherche d'emploi, etc.). Comme pour le précédent corpus, l'emploi généralisé du *plain language* sur le site britannique peut également expliquer ces disparités.

Résultats et discussions

Le premier ensemble de corpus dit classique nous a permis d'observer les spécificités terminologiques et phraséologiques du discours, en étudiant par exemple les groupes nominaux complexes se référant aux aides sociales comme « Allocation de soutien familial », ou « *Personal Independence Payment* ». Ces groupes nominaux complexes sont tous formés sur le même modèle. Ils sont composés d'un nom « noyau », synonyme ou quasi- synonyme d'aide financière : *allocation, prestation, complément, prime, aide* en français ou *payment, support, credit, benefits,* en anglais. Ce nom est toujours suivi d'un complément : adjectif ou groupe prépositionnel qui indique le plus souvent le domaine des aides : *logement, enfant, rentrée scolaire*, etc. Nous avons identifié des tendances dans la formation de ces groupes nominaux, que nous avons formalisés par le schéma suivant :

[noyau] [préposition] [complément1] [complément2] [complément3] […]

La formalisation de ces groupes nominaux nous a permis de les étudier de façon systématique dans le premier ensemble de corpus.

Les recherches dans ce premier ensemble de corpus ont également permis d'identifier une prévalence de l'utilisation des pronoms « you » et leur équivalent « vous », qui place l'usager en tant que sujet principal de ce discours. Ce pronom est également présent dans des unités phraséologiques comme l'unité « si vous » ou son équivalent en anglais britannique « if you ». Ces résultats pourraient être en lien avec le replacement de la responsabilité sur le citoyen qui doit être « entrepreneur » de ses droits (Lochak, 2022). Elle donne également lieu à des schémas phraséologiques comme le suivant : « [*Si vous êtes] [situation administrative]* », qui pourrait être mis en lien avec des recherches sur la montée en conditionnalité de l'accès aux droits (Chelle, 2019).

Le second ensemble de corpus issu du parcours utilisateur permet d'envisager le discours dans son interactivité vis-à-vis de la page web, et de mettre en perspective les observations faites à partir du premier ensemble de corpus. En effet, il permet d'observer le rôle que jouent les hyperliens, les formulaires et les menus déroulants sur la page. Nous avons par exemple remarqué que les définitions de termes et groupes nominaux complexes correspondaient à des parties déroulantes sur la page, mais aussi à des hyperliens : des éléments interactifs avec lesquels il est nécessaire d'interagir pour obtenir une information. Ces résultats peuvent également être mis en lien avec une montée en conditionnalité de l'accès aux droits liée à l'utilisation du numérique (Mazet, 2019).

L'association de l'étude des deux ensembles de corpus permet de nourrir nos analyses et d'enrichir nos résultats. Elle offre trois pistes de réflexion. Tout d'abord, elle nous pousse à interroger l'objet d'étude qu'est le corpus et sa place dans les transformations numériques du discours. Dans un second temps, elle permet de s'intéresser à la place des discours spécialisés dans les transformations numériques. Enfin, elle nous permet d'explorer la manière dont nous pouvons en tirer des connaissances utiles pour les formations de rédacteurs et traducteurs spécialisés, sur le discours administratif numérique et plus largement sur les discours spécialisés numériques.

Références bibliographiques

Adam, J.-M. (1997). Genres, textes, discours : Pour une reconception linguistique du concept de genre. *Revue belge de philologie et d'histoire*, *75*(3), 665-681. https://doi.org/10.3406/rbph.1997.4188

Authier, J. (1982). La mise en scène de la communication dans des discours de vulgarisation scientifique. *Langue Française*, *53*, 34-47.

Bakhtine, M. (1984). Les genres du discours. In *Esthétique de ma création verbale*, *265*, 308.

Bathia, V. (2004). *Worlds of Written Discourse: A Genre-Based View.* (Vol. 148). Bloomsbury Academic.

Cacchiani, S. (2018). Représenter et communiquer les connaissances spécialisées sur Copyright dans des références juridiques de consultation rapide et sur des plateformes en ligne institutionnelles. *Éla. Études de linguistique appliquée*, *192*(4), 505-521. https://doi.org/10.3917/ela.192.0505

Chelle, E. (2019). *Gouverner les pauvres : Politiques sociales et administration du mérite.* (Presses universitaires de Rennes.).

Delavigne, V. (2022). La notion de domaine en question À propos de l'environnement. *Neologica 2022, n° 16. Néologie et environnement*, 27-59. https://doi.org/10.48611/isbn.978-2-406-13219-6.p.0027

François, D. (2014). Le droit à l'information selon la DILA. *Documentaliste : Sciences de l'Information*, *51*(4), 32-33. https://doi.org/10.3917/docsi.514.0032

Lerat, P. (1995). *Les langues spécialisées*. Presses Universitaires de France.

Lochak, D. (2022). Le sans contact, nouvelle norme du service public : *Plein droit, n° 134(3)*, 3- 6. https://doi.org/10.3917/pld.134.0005

Maingueneau, D. (2016). L'ethos discursif et le défi du Web. *Itinéraires*, *2015-3*, 2015-2018. https://doi.org/10.4000/itineraires.3000

Mazet, P. (2021). *Les conditionnalités implicites de l'accès aux droits à l'ère numérique*. https://shs.hal.science/halshs-03218656

Paveau, M.-A. (2015). Ce qui s'écrit dans les univers numériques. *Itinéraires*, *2014-1*. https://doi.org/10.4000/itineraires.2313

Paveau, M.-A. (2017). *L'analyse du discours numérique. Dictionnaire des formes et des pratiques*. Hermann. https://journals.openedition.org/lectures/24355

Simon, J. (Éd.). (2018). *Le discours hypertextualisé : Espaces énonciatifs mosaïques*. Presses universitaires de Franche-Comté. https://doi.org/10.4000/books.pufc.40815

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Souchier, E., Candel, É., & Gomez-Mejia, G. (2019). *Le numérique comme écriture. Théories et méthodes d'analyse*. Armand Colin.

Swales, J. M. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge University Press.

Swales, J. M. (2004). Research genres: Explorations and applications. Cambridge University Press.

Comparaison diachronique de motifs récurrents dans deux encyclopédies

Alice Brenon[1, 2], Denis Vigier[1], Ludovic Moncla[2] et Frédérique Laforest[2]
[1] ICAR, CNRS, UMR 5191, F-69342 [2] Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR 5205, F-69621
alice.brenon@insa-lyon.fr

Introduction

En choisissant le titre *Encyclopédie ou dictionnaire raisonné des sciences, des arts et des métiers* (*EDdA* dans ce qui suit), Diderot et d'Alembert inscrivent leur œuvre dans une double lignée. Ils reconnaissent d'abord l'héritage de la *Cyclopædia* de Chambers parue en 1728 à Londres et dont l'*EDdA* ne devait être initialement qu'une traduction (Kafker & Loveland, 2016, p. 111 et seq.). Ils reconnaissent ensuite celle des dictionnaires universels comme celui de Furetière qui inaugura cette lignée en 1690 ou comme le *Dictionnaire Universel François et Latin*, dit « *de Trévoux* » qui fut au long de ses éditions successives un grand rival de l'*EDdA* dont il critique le projet philosophique. On remarque toutefois qu'à la différence de la *Cyclopædia* qui arbore l'expression *Universal Dictionary* dans son sous-titre, les encyclopédistes français lui préfèrent l'adjectif *raisonné*, reflet de l'approche progressiste de ses auteurs, emblématique du siècle des Lumières.

Parmi les domaines que traite l'*EDdA*, la géographie figure en bonne position : elle est le domaine qui comporte le plus d'articles (environ 20% du total). Et pour cause, au XVIII^{ème} siècle, grand siècle d'exploration, la géographie est une science à la mode, encore largement rattachée aux mathématiques (ce qui est visible jusque dans le « Système figuré des connoissances humaines », placé en tête de l'ouvrage).

Plusieurs études dont Vigier et al. (2022) ont contribué à montrer qu'à cette époque les articles encyclopédiques traitant de géographie se différencient peu des autres productions du même domaine, qu'elles soient issues d'ouvrages généralistes comme les dictionnaires universels ou plus spécialisés comme par exemple *Le Grand Dictionnaire géographique* de Bruzen de La Martinière (1726-1739) ou le *Dictionnaire géographique-portatif* de Vosgien (1747).

Tout porte à croire en revanche que le profil générique et discursif des articles de géographie a changé sensiblement au cours du XIX^{ème} siècle, sous l'influence directe d'une évolution majeure des discours scientifiques de tous les domaines à cette époque. Il a en effet été montré (Jacquet-Pfau, 2022) que ce siècle — et surtout sa seconde moitié — a connu une « disciplinarisation » des savoirs du fait de l'importance grandissante des universités dans leur production. Ce fait peut s'observer par exemple dans la composition du groupe des rédacteurs de *La Grande Encyclopédie ou Inventaire raisonné des Sciences, des Lettres et des Arts* (ci-après *LGE*), une œuvre majeure de la fin du XIX^{ème} siècle (la parution débute en 1885 pour prendre fin en 1902) qui s'inscrit dans la tradition de l'*EDdA* (Jacquet-Pfau, 2015, p. 88 et seq.), ce qui est reflété jusque dans son titre complet.

Dans cette communication, qui s'inscrit dans le contexte du projet GEODE[1], nous voulons donc explorer l'hypothèse selon laquelle les articles encyclopédiques de géographie diffèrent

sensiblement à la fin du XIX^{ème} siècle des articles du même domaine publiés un siècle plus tôt dans l'*EDdA*.

Corpus d'étude

Pour tester cette hypothèse, nous nous proposons de comparer l'*EDdA* à cette seconde encyclopédie plus récente qu'est *LGE*. Nous utiliserons pour les articles de l'*EDdA* la version publiée par l'ARTFL[2] et diffusée sur son site web (Morrissey & Roe, 2022). Cette partie de notre corpus représente 17 tomes (les 17 volumes de texte de l'*EDdA*, auxquels s'ajoutent 11 volumes de planches que nous laissons de côté pour notre étude) et comprend au total 19,6 millions de mots dans 74 190 articles.

La version de *LGE* que nous verserons à notre corpus est issue des résultats du projet DISCO-

LGE[23] (Vigier & Brenon, 2021). La segmentation de ses 31 tomes est encore imparfaite mais elle comprend dans sa version actuelle 134 820 articles pour 56,6 millions de mots. Elle ne comporte pas de métadonnées sinon le titre des entrées, le tome dont elles proviennent et leur position dans la séquence des articles. Les logiciels développés pour réaliser ce découpage sont disponibles publiquement[24].

Afin de ne mesurer que les différences dans la manière d'écrire et pas dans les sujets traités, relevant davantage de la variation du périmètre même du domaine géographique que de changements stylistiques, nous nous concentrerons sur des articles décrivant les mêmes objets dans l'*EDdA* et dans *LGE*. Dans ce but, nous avons constitué un sous-corpus de notre corpus d'étude initial, que nous nommerons *Corpus Parallèle*, qui comprend les articles communs à l'*EDdA* et à *LGE*, c'est à dire les articles ayant la même vedette, unique à la fois dans l'*EDdA* et dans *LGE*, afin de limiter les ambiguïtés et donc les erreurs d'appariement. Ce corpus parallèle contient 7 215 paires d'articles de chacune des deux encyclopédies (soit 14 430 articles au total) pour un total de 13,2 millions de mots environ, avec un déséquilibre en faveur de *LGE* légèrement moins important que pour le corpus complet (8,7 millions contre 4,5 millions).

Transversalement, pour pouvoir également tenir compte de la variation diachronique des frontières entre domaines dans les deux encyclopédies, un deuxième sous-corpus a été défini en intégrant tous les articles portant sur une même thématique mais des objets potentiellement différents. Vigier et al. (2020, p. 7) ont repéré des motifs syntaxiques apparaissant préférentiellement en début d'articles et permettant d'identifier des « mots classifieurs ». À leur suite, nous avons donc constitué un corpus des hydronymes en choisissant les articles pour lesquels ce mot appartient au lexique des cours et plans d'eau (rivière, fleuve, lac, etc.). Ce deuxième sous-corpus contient 1354 articles de l'*EDdA* et 1917 de *LGE*, représentant 99k mots pour l'*EDdA* et 261k pour *LGE*.

Méthodologie

Afin de pouvoir contraster les articles de géographie avec ceux des autres domaines, il est nécessaire d'avoir un étiquetage des domaines de connaissances des articles. Un modèle de classification entraîné de manière supervisée sur l'*EDdA* pour un ensemble de domaines

[23] . https://www.collexpersee.eu/projet/disco-lge/

[24] . https://gitlab.huma-num.fr/disco-lge

communs aux deux encyclopédies et s'appuyant sur un modèle de langue pré-entraîné (Brenon et al., 2022, p. 6) pour classer les articles a été appliqué sur *LGE*. Le croisement de cette information avec les deux sous-corpus permettra à la fois d'interroger la pertinence de leur définition et les évolutions propres qu'a pu subir chaque domaine : si tous les hydronymes devraient relever de la géographie, le corpus parallèle contient probablement en revanche un certain nombre d'homonymes dont l'appariemment accidentel était indésirable mais sans doute aussi des examples d'objets dont une discipline s'est emparée au détriment d'une autre entre les deux époques.

La possibilité de partitionner le corpus d'étude complet par œuvre et par domaine ouvre la perspective de calculs de spécificités à l'aide d'outils comme TXM (Heiden, 2010). L'étude des coocurrences présentes dans les différentes parties du corpus révélera si les motifs présents dans l'*EDdA* le sont encore dans *LGE*.

Mais, faisant l'hypothèse que la « disciplinarisation » dont il est question plus haut s'accompagne d'une normalisation de la production écrite de chaque science passant possiblement par des routines discursives spécifiques, nous utiliserons également le Lexicoscope (Kraif, 2016) pour détecter des motifs syntaxiques plus profonds et peu visibles en restant au niveau de la surface des phrases.

Pour travailler ainsi au niveau de la syntaxe aussi bien que des parties de discours, nous avons annoté le corpus avec la librairie Python *Stanza* (Qi et al., 2020) pour le modèle French-GSD[25] utilisant le jeu d'étiquettes des *Universal Dependencies* (Marneffe et al., 2021).

Premiers résultats

Une étude préliminaire du corpus des hydronymes a permis de mettre en évidence dans *LGE* des tournures de phrases déjà apparue l'*EDdA*, comme le motif suivant utilisé fréquemment dans les articles : « ... prend sa source [...] et se jette [...] ».

Une différence mineure a cependant été constatée dans les réalisations de ce motif entre les deux œuvres : le sujet en est plus fréquemment le pronom relatif «qui» dans l'*EDdA*, ce qui permet de l'introduire dans la phrase nominale, souvent unique, qui définit le cours d'eau, alors qu'il fait davantage l'objet dans *LGE* d'une phrase séparée où il commence par le pronom personnel « il » ou «elle». De plus, certains articles dans lesquels le motif précédent n'a pas été trouvé possèdent toutefois une structure voisine où la description du cours d'eau a bien lieu de sa source à son embouchure mais au travers d'une chaîne de coréférence plus complexe pouvant occuper tout un paragraphe avec une plus grande variabilité lexicale au niveau des verbes employés.

Ces premiers résultats suggèrent qu'une évolution a eu lieu et amènent à s'interroger sur l'existence possible de tournures de phrases propres à l'une ou l'autre des deux œuvres du corpus, ou sur ce qui pourrait distinguer leur emploi dans celles qui leur sont communes. Sur le plan quantitatif il serait également intéressant de comprendre comment sont redistribués les volumes de mots entre les domaines de connaissance. Le phénomène se limite-t-il aux hydronymes ou concerne-t-il toute la géographie?

[25] . https://universaldependencies.org/treebanks/fr_gsd/

Remerciements

Les auteurs remercient le LABEX ASLAN (ANR-10-LABX-0081) de l'Université de Lyon pour son soutien financier dans le cadre du programme français "Investissements d'Avenir" géré par l'Agence Nationale de la Recherche (ANR).

Références bibliographiques

Brenon, A., Moncla, L., & Mcdonough, K. (2022). Classifying encyclopedia articles : Comparing machine and deep learning methods and exploring their predictions. Data and Knowledge Engineering, 102098. https://doi.org/10.1016/j.datak.2022.102098

Heiden, S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In R. Otoguro, K. Yoshimoto, K. Ishikawa, H. Umemoto, & Y. Harada (Eds.), 24th Pacific Asia Conference on Language, Information and Computation (Vol. 2, p. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University. https://halshs.archives-ouvertes.fr/halshs-00549764

Jacquet-Pfau, C. (2022). Actualiser l'Encyclopédie de Diderot et d'Alembert au XIXe siècle : Le Grand Dictionnaire universel (1866-1890) et La Grande Encyclopédie (1885-1902) : Langue Française, N° 214(2), 95–109. https://doi.org/10.3917/lf.214.0095

Jacquet-Pfau, C. (2015). Élaboration et destinée d'une encyclopédie la fin du XIXe siècle : Les trente- et-un volumes de la grande encyclopédie, inventaire raisonné des sciences, des lettres et des arts par une société de savants et de gens de lettres. Éla. Études de Linguistique Appliquée, 177, 85–100. https://doi.org/10.3917/ela.177.0085

Kafker, F. A., & Loveland, J. (2016). André-François Le Breton, initiateur et libraire en chef de l'Encyclopédie. Recherches Sur Diderot Et Sur l'Encyclopédie, 51, 107–125. https://doi.org/10.4000/rde.5390

Kraif, O. (2016). Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. Cahiers de Lexicologie, 1(108), 91–106. https://doi.org/10.15122/isbn.978-2-406-06281-3.p.0091

Marneffe, M.-C. de, Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. Computatio- nal Linguistics, 47(2), 255–308. https ://doi.org/10.1162/coli$_{a0}$0402 Morrissey, R., & Roe, G. (2022). Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, etc. University of Chicago : ARTFL Encyclopédie Project. https://encyclopedie.uchicago.edu/

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python natural language pro- cessing toolkit for many human languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations. https://nlp.stanford.edu/pubs/qi2020stanza.pdf Vigier, D., & Brenon, A. (2021). La Grande Encyclopédie ou Inventaire raisonné des Sciences, des Lettres et des Arts. 264458212, 185163346, 398759061, 1371168210. https://doi.org/10.34847/NKL.74EB1XFD

Vigier, D., Moncla, L., Brenon, A., Mcdonough, K., & Joliveau, T. (2020). Classification des entités nommées dans l'Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers par une société de gens de lettres (1751-1772). In Actes du 7ème Congrès Mondial de

Linguistique Française, Jul 2020, Montpellier, France. https://doi.org/10.1051/shsconf/20207811008

Vigier, D., Moncla, L., Lefort, I., Joliveau, T., & McDonough, K. (2022). Les articles de géographie dans le Dictionnaire Universel de Trévoux et l'Encyclopédie de Diderot et d'Alembert : Langue Française, N° 214(2), 59–80. https://doi.org/10.3917/lf.214.0059

Quant à lui/eux versus *lui/eux*: Influence of register and syntactic complexity on their alternation and syntactic position

Jorina Brysbaert [1], Karen Lahousse [1] and Benedikt Szmrecsanyi [1]
[1] KU Leuven
jorina.brysbaert@kuleuven.be, karen.lahousse@kuleuven.be, benedikt.szmrecsanyi@kuleuven.be

Introduction

In this talk, we present a corpus analysis of two competing, syntactically mobile markers for contrastive topics in different registers of French: the emphatic pronoun *lui/eux* 'him/them' (E-Pro) (1) and the emphatic pronoun *lui/eux* 'him/them' introduced by the preposition *quant à* 'as for' (*quant à* E-Pro) (2).

(1) *Alice adore les chats. Martin **[lui]** préfère **[lui]** les chiens **[lui]**.*
Alice loves cats. Martin **[him]** prefers **[him]** dogs **[him]**. (lit.)

(2) *Alice adore les chats. Martin **[quant à lui]** préfère **[quant à lui]** les chiens **[quant à lui]**.*
Alice loves cats. Martin **[as for him]** prefers **[as for him]** dogs **[as for him]**. (lit.)

Our goal is to show that both the choice for a specific marker (E-Pro vs. *quant à* E-Pro) and the choice for a specific syntactic position (between subject & verb vs. inside/after verb phrase) are influenced by the syntactic complexity of the subject and the register.

Previous research on these markers for contrastive topics is rather scarce and is not based on extensive corpus analyses (on E-Pros: Caddéo, 2004, 2006, 2008; Cappeau, 1999; Carton, 2009; Nølke, 1997; Rocquet, 2014; on *quant à* E-Pros: Anscombre, 2006; Choi-Jonin, 2003; Debaisieux, 2001; Fløttum, 1999, 2003; Lagae, 2003, 2007, 2011; Rocquet, 2014). Moreover, to the best of our knowledge, the two markers have never been compared.

Corpora and methodology

We extracted occurrences of E-Pros (4965 examples in total) and *quant à* E-Pros (1804 examples in total) from six corpora, using AntConc (Anthony, 2018). The properties of the corpora are summarized in table 1. Three of them consist of written data: *Le Monde* (LM) (SA Le Monde, 1999), which contains articles from the national quality newspaper *Le Monde*; *Est Républicain* (ER) (Gaiffe et al., 2018), which consists of articles from the regional newspapers *L'Est Républicain* and *Vosges Matin*; and *Yahoo Contrastive Corpus of Questions and Answers* (YCC), containing texts from the former online discussion platform *Yahoo! Answers* (De Smet, 2009). The other three corpora contain spoken French and consist mainly of interviews between an interviewer and one or more interviewees who are not familiar with each other: *Corpus Oral de Français de Suisse Romande* (OFROM) (Avanzi et al., 2012-2020), *Corpus de Français Parlé Parisien des années 2000* (CFPP) (Branca-Rosoff et al., 2011, 2012) and *Corpus de Référence du Français Parlé* (CRFP) (Équipe DELIC, 2004).

	LM	ER	YCC	OFROM	CFPP	CRFP
Mode	written	written	written	spoken	spoken	spoken
Level of formality	formal	semi-formal	informal	semi-formal	semi-formal	semi-formal
Type	national newspaper	regional newspaper	online discussion platform	interview	interview	interview & public speech
Year	1998	2010-2011	2006-2009	2012-2020	2005-2019	1998-2004
Region	France	north-east of France	unknown	Switzerland	Paris & suburbs	38 towns across France
Number of words	25.7 million	74 million	6.1 million	1.01 million	0.7 million	0.44 million

table 4. : Properties of the six corpora

The six corpora are representative of different 'registers' of French, defined as the intersection of mode (spoken versus written) and formality (formal versus informal), based on the continuum between 'language of distance' and 'language of immediacy' developed by Koch & Oesterreicher (2007). Since the spoken corpora (between 0.44 and 1.01 million words) are much smaller than the written corpora (between 6.1 and 74 million words), we will group together the results for the three spoken corpora.

Results

Choice for marker

Our data show that E-Pros are in general more frequent than *quant à* E-Pros (see table 2), but there is also an interesting **register effect**, which is statistically significant (χ^2 = 326.29, N = 6769, df = 3, $p < 0.001$, Cramer's V = 0.22). In the informal written YCC, a less complex register, the proportion of *quant à* E-Pros is significantly lower (10%) than in the semi-formal written ER (34%), a more complex register. *Quant à* E-Pros are even completely absent from our spoken corpora (0%), which is in line with Debaisieux (2001), who finds only 23 examples of *quant à X* 'as for X' (with *X* being any possible element) in her spoken corpus of 1.250.000 words (18,4 example/million words). Our results thus provide evidence for the idea that more explicit markers are more frequent in more complex registers (cf. Szmrecsanyi & Engel 2023 on English and Dutch).

	LM	ER	YCC	SPOKEN
E-Pros	85% (1789)	66% (2812)	90% (282)	100% (82)
Quant à E-Pros	15% (326)	34% (1446)	10% (32)	0% (0)
Total	100% (2115)	100% (4258)	100% (314)	100% (82)

table 5. : E-Pros vs. *quant à* E-Pros: frequency per register

However, the result for the LM corpus does not entirely fit into the general tendency. Following Szmrecsanyi & Engel (2023), this corpus represents in theory the most complex register, since it consists of formal written data from a high quality newspaper. Hence, one might expect it to contain the highest number of *quant à* E-Pros, but this is definitely not the case. Their proportion only amounts to 15% in LM, in contrast to 34% in the semi-formal newspaper corpus ER. This difference is hard to explain, but we hypothesize that the journalists from LM might resort more frequently to other marking strategies, which also have a high degree of formal or functional explicitness, such as contrastive adverbs (e.g. *en*

revanche 'on the other hand'), *pour sa part* 'for his/her/its part' or *de son côté* 'for his/her/its side'.

In addition, in cases with a lexical NP subject (as *Martin* in (1)-(2)), the alternation between E-Pros and *quant à* E-Pros is influenced by the **syntactic complexity of the subject**, defined as post-modification of the subject (e.g. by means of an apposition: *Martin, son frère, préfère les chiens.* 'Martin, her brother, prefers dogs.') (see table 3).

	LM		**ER**		**YCC**		**SPOKEN**	
	E-Pros	*Quant à* E-Pros	E-Pros	*Quant à* E-Pros	E-Pros	*Quant à* E-Pros	E-Pros	*Quant à* E-Pros
Subject without post-modification	70% (1135)	51% (165)	70% (1754)	55% (785)	85% (140)	81% (26)	87% (20)	0
Subject with post-modification	30% (478)	49% (161)	30% (744)	45% (637)	15% (25)	19% (6)	13% (3)	0
Total	100% (1613)	100% (326)	100% (2498)	100% (1422)	100% (165)	100% (32)	100% (23)	0

table 6. : E-Pros vs. *quant à* E-Pros: syntactic complexity of subject per register

In the two journalistic written corpora (LM and ER), the proportion of complex post-modified subjects is significantly higher with the more explicit *quant à* E-Pros than with the less explicit E-Pros (LM: $\chi^2 = 47.89$, $N = 1939$, $df = 1$, $p < 0.001$, Cramer's $V = 0.16$; ER: $\chi^2 = 89.50$, $N = 3920$, $df = 1$, $p < 0.001$, Cramer's $V = 0.15$). This finding is in line with the Complexity Principle, as described by Rohdenburg (2020), according to which more explicit items are more frequent in more complex grammatical contexts. In other words, more complex post-modified subjects promote the use of the more explicit *quant à* E-Pros, as shown in (3).

(3) *Sourire aux lèvres, M. Sarkozy a ostensiblement serré longuement la main du Premier ministre sortant à son départ de l'Elysée peu après 19 h 30. […] Le numéro deux du gouvernement, Jean-Louis Borloo, qui convoitait Matignon, a été reçu* ***quant à lui*** *vers 17 h 30 par le président.*
With a smile on his lips, Mr. Sarkozy ostensibly shook hands with the outgoing Prime Minister for a long time at his departure from the Elysee Palace shortly after 7:30 p.m. […] The number two of the government, Jean-Louis Borloo, who coveted Matignon, was received **as for him** around 5:30 p.m. by the president. (ER)

By contrast, in the informal written YCC, there is no statistically significant difference between the E-Pros and the *quant à* E-Pros in terms of syntactic complexity of their subject ($\chi^2 = 0.26$, $N = 197$, $df = 1$, $p = 0.61$). Both markers combine most frequently with a non-post-modified lexical NP (85% and 81%). This is probably due to the overall lower frequency of post-modified subjects in the YCC, and to the fact that the post-modified subjects in this corpus are overall less syntactically complex than those in LM and ER, as illustrated in (4) (compared to (3)).

(4) *Mon beau papa travaille dans une société qui en vends, il est VRP. Son fils (mon chéri quoi!)* ***lui*** *travaille sur la fabrication des appareils de radiographie classiques et numériques dans les hôpitaux et les cabinets privés.*
My father-in-law works in a company that sells them, he is a sales representative. His son (my darling thus!) **him** works on the production of conventional and digital X-ray devices in hospitals and private practices (YCC)

Choice for syntactic position

As becomes clear from table 4, **register** also has an influence on the syntactic position of E-Pros and *quant à* E-Pros: both markers appear significantly more often after the finite verb (i.e. within or after the VP) in the two journalistic written corpora (LM and ER) than in the informal written YCC and the spoken corpora (E-Pros: $\chi^2 = 126.57$, $N = 4965$, $df = 1$, $p < 0.001$, Cramer's $V = 0.16$; *quant à* E-Pros: $\chi^2 = 11.76$, $N = 1804$, $df = 1$, $p < 0.001$, Cramer's $V = 0.08$).

	LM		**ER**		**YCC**		**SPOKEN**	
	E-Pros	*Quant à* E-Pros	E-Pros	*Quant à* E-Pros	E-Pros	*Quant à* E-Pros	E-Pros	*Quant à* E-Pros
Between subject and finite verb	61% (1093)	25% (82)	54% (1519)	28% (401)	83% (235)	56% (18)	82% (67)	0
After finite verb	39% (696)	75% (244)	46% (1293)	72% (1045)	17% (47)	44% (14)	18% (15)	0
Total	100% (1789)	100% (326)	100% (2812)	100% (1446)	100% (282)	100% (32)	100% (82)	0

table 7. : Syntactic position of E-Pros vs. *quant à* E-Pros: influence of register

Moreover, there is a significant effect of the **syntactic complexity of the subject** (defined in terms of post-modification) on the syntactic position of both markers. Post-modified subjects boost the use of the position after finite verb, in order to avoid adding extra lexical material – in the form of a (*quant à*) E-Pro – between the head of the subject and the finite verb. However, the impact of post-modification is higher with the E-Pros (table 5) than with the *quant à* E-Pros (table 6), which occur overall most often after the finite verb, due to their 'heavier' syntactic weight.

	LM		**ER**		**YCC**		**SPOKEN**	
	Post-modified	Not post-modified	Post-modified	Not post-modified	Post-modified	Not post-modified	Post-modified	Not post-modified
Between subject and finite verb	28% (137)	74% (956)	27% (206)	64% (1313)	67% (22)	86% (213)	75% (3)	82% (64)
After finite verb	72% (355)	26% (341)	73% (545)	36% (748)	33% (11)	14% (36)	25% (1)	18% (14)
Total	100% (492)	100% (1297)	100% (751)	100% (2061)	100% (33)	100% (249)	100% (4)	100% (78)

table 8. : Syntactic position of E-Pros: influence of syntactic complexity of S

	LM		**ER**		**YCC**	
	Post-modified	Not post-modified	Post-modified	Not post-modified	Post-modified	Not post-modified
Between subject and finite verb	10% (16)	40% (66)	16% (101)	37% (300)	17% (1)	65% (17)
After finite verb	90% (145)	60% (99)	84% (539)	63% (506)	83% (5)	35% (9)
Total	100% (161)	100% (165)	100% (640)	100% (806)	100% (6)	100% (26)

table 9. : Syntactic position of *quant à* E-Pros: influence of syntactic complexity of S

This combined effect of register and syntactic complexity of the subject on the syntactic position of (*quant à*) E-Pros is further confirmed by a conditional inference tree analysis (figure 1 for E-Pros and figure 2 for *quant à* E-Pros) (Hothorn et al. 2006). For both markers, the first split (Node 1) is based on the presence vs. absence of post-modification, which

indicates that this is the most powerful factor in determining the choice for a certain syntactic position. When the subject is post-modified (branch to the right), the marker occurs less often between the subject and the finite verb than when it is not post-modified (branch to the left).

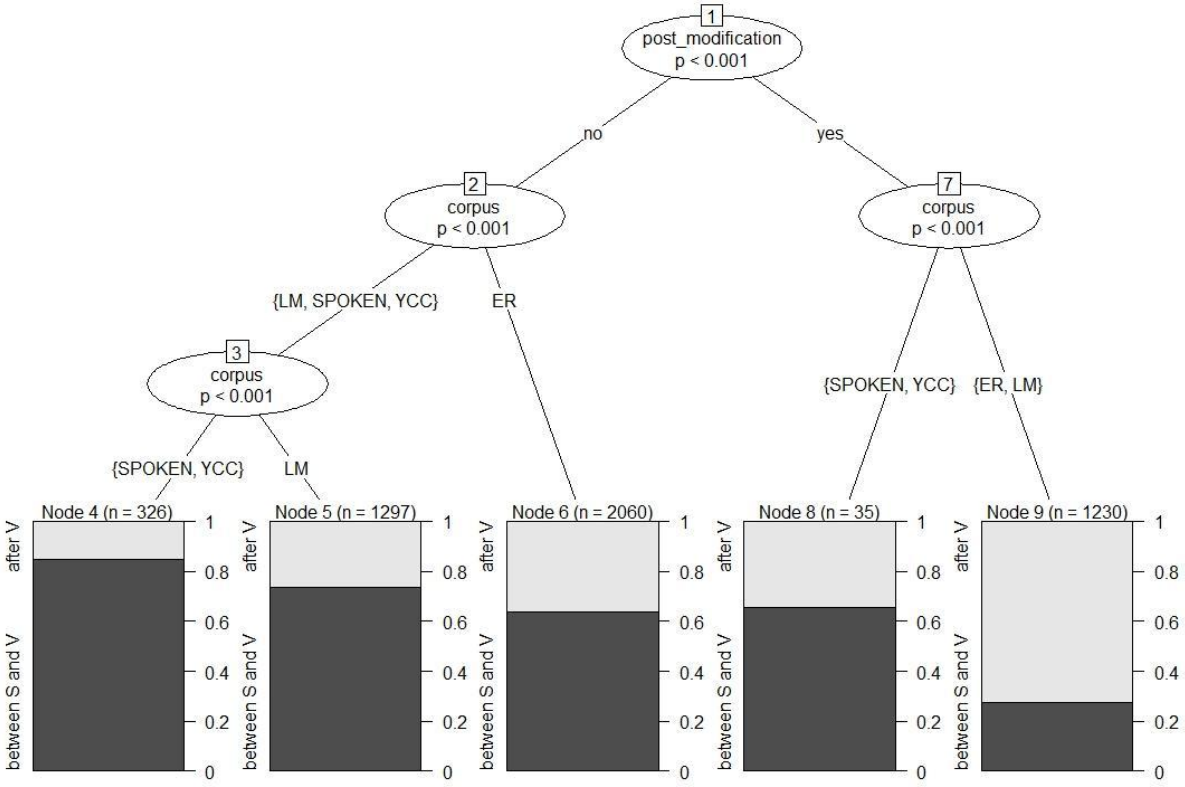


figure . 64 Conditional inference tree: factors influencing the syntactic position of E-Pros

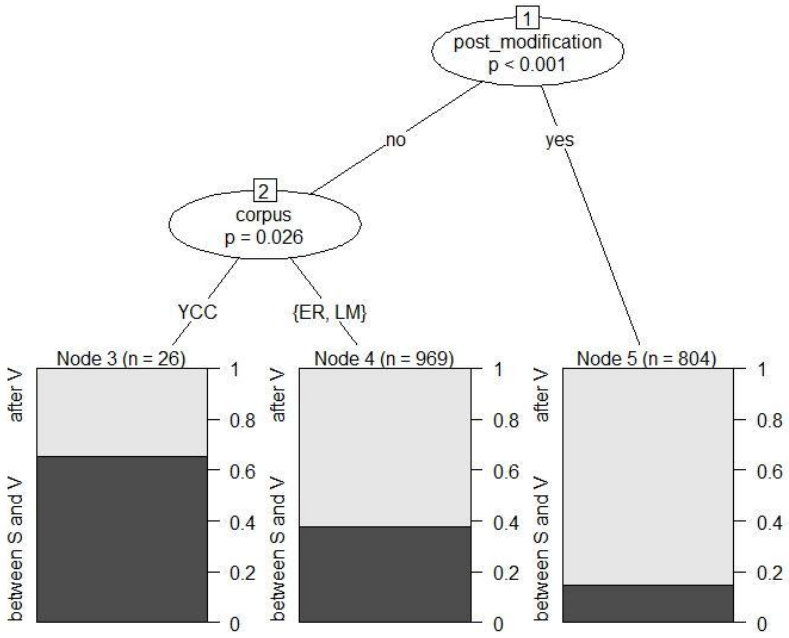


figure . 65 Conditional inference tree: factors influencing the syntactic position of *quant à* E-Pros

References

Anscombre, J.-C. (2006). Les locutions *quant à*, *pour ce qui est de*, *en ce qui concerne* : chronique d'un discours annoncé. *Modèles Linguistiques*, 54, 155-169.

Anthony, L. (2018). *AntConc (Version 3.5.7) [computer software]*. Waseda University. Available from https://www.laurenceanthony.net/software.

Avanzi, M., Béguelin, M.-J., Corminboeuf, G., Diémoz, F., Johnsen, L.A. (2012-2020). *Corpus OFROM – Corpus oral de français de Suisse romande [corpus]*. Université de Neuchâtel.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., Pires, M. (2011). Constitution et exploitation d'un corpus de français parlé parisien. *Corpus*, 10, 81-98.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., Pires, M. (2012). Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000). http://cfpp2000.univ-paris3.fr/CFPP2000.pdf

Caddéo, S. (2004). Lui, le propriétaire, le propriétaire, lui: Deux constructions bien distinctes. *Recherches sur le Français Parlé*, 18, 145-161.

Caddéo, S. (2006). Apposition ? À la limite ! *L'Information Grammaticale*, 109, 34-37.

Caddéo, S. (2008). L'apposition : une construction multiforme. *Travaux de Linguistique*, 57(2), 63-72.

Cappeau, P. (1999). Sujets éloignés. Esquisse d'une caractérisation des sujets lexicaux séparés de leur verbe. *Recherches sur le Français Parlé*, 15, 199-231.

Carton, F. (2009). Étude prosodique d'un cas de détachement. Les pronoms personnels pseudo-disjoints dans un corpus de presse parlée en français. In D. Apothéloz, B. Combettes, F. Neveu (Eds.), *Les linguistiques du détachement. Actes du colloque international de Nancy (7-9 juin 2006)* (pp. 173-187). Peter Lang.

Choi-Jonin, I. (2003). Ordre syntaxique et ordre référentiel: Emplois de la locution prépositive *quant à*. In B. Combettes, C. Schnedecker, A. Theissen (Eds.), *Ordre et distinction dans la langue et le discours* (pp. 133-147). Honoré Champion.

De Smet, H. (2009). *Yahoo contrastive corpus of questions and answers [corpus]*. Department of Linguistics, KU Leuven.

Debaisieux, J.-M. (2001). Contraintes syntaxiques et discursives des emplois de *quant à* et *en ce qui concerne* en français parlé. *Cahiers de Praxématique*, 37, 125-146.

Engel, A. (2022). The register-specificity of probabilistic grammars in English and Dutch. Combining corpus analysis and experimentation. PhD dissertation. KU Leuven.

Équipe DELIC. (2004). Présentation du Corpus de référence du français parlé. *Recherches sur le Français Parlé*, 18, 11-42.

Fløttum, K. (1999). *Quant à* - thématisateur et focalisateur. In C. Guimier (Ed.), *La thématisation dans les langues* (pp. 135-149). Peter Lang.

Fløttum, K. (2003). À propos de 'quant à' et 'en ce qui concerne'. In B. Combettes, C. Schnedecker, A. Theissen (Eds.), *Ordre et distinction dans la langue et le discours* (pp. 185-202). Honoré Champion.

Gaiffe, B., Nehbi, K., Tonnelier, M. (2018). *Corpus journalistique issu de l'Est Républicain [corpus]*. In (Version 2) Analyse et traitement informatique de la langue française [ATILF]. Available from ORTOLANG, https://hdl.handle.net/11403/est_republicain/v2.

Hothorn, T., Hornik, K., Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistic*, 15(3), 651-674.

Koch, P., Oesterreicher, W. (2007). Schriftlichkeit und kommunikative Distanz. *Zeitschrift für Germanistische Linguistik*, 35(3), 346-375.

Lagae, V. (2003). *Quant aux livres / De livre, il n'en a lu aucun*. Étude syntaxique de deux constructions détachées. *Lingvisticae Investigationes*, 26(2), 235-258.

Lagae, V. (2007). Left-detachment and topic-marking in French: The case of *quant à* and *en fait de*. *Folia Linguistica*, 41(3-4), 327-355.

Lagae, V. (2011). Le paradigme des marqueurs thématiques en français : essai de typologie. In E. Comes, S. Miculescu (Eds.), *La construction d'un paradigme. Actes du XVIIIe séminaire de didactique universitaire Constanta 2010* (pp. 53-73). Editura Echinox.

Nølke, H. (1997). Anaphoricité et focalisation: Le cas du pronom personnel disjoint. In W. De Mulder, L. Tasmowski-de Ryck, C. Vetters (Eds.), *Relations anaphoriques et (in)cohérence* (pp. 55-67). Brill/Rodopi.

Rocquet, A. (2014). The discourse-marking effect of strong pronoun doubling in French. *Phrasis: Studies in Language and Literature*, 50, 95-112.

Rohdenburg, G. (2020). The Complexity Principle at work with rival prepositions. *English Language & Linguistics*, 24(4), 769-800.

SA Le Monde. (1999). *Le Monde sur CD-ROM: CEDROM-SNI [corpus]*.

Szmrecsanyi, B., Engel, A. 2023. A variationist perspective on the comparative complexity of four registers at the intersection of mode and formality. *Corpus Linguistics and Linguistic Theory*, 19(1), 79-113.

Annotation lexicale et pragmatique de termes médicaux et leurs reformulations

Ioana Buhnila [1]
[1] Laboratoire LiLPa UR 1339, Université de Strasbourg
ioana.buhnila@etu.unistra.fr

Introduction

Les connaissances scientifiques du domaine médical se développent à un rythme accéléré pour faire face aux défis sanitaires constants et évolutifs. Pourtant, les notions médicales restent tout aussi opaques pour le grand public, non-spécialiste du domaine de la médecine. Pour comprendre les informations médicales transmises, le grand public a besoin de *variantes simplifiées* de ces termes médicaux. Dans ce sens, nous nous intéressons aux *reformulations médicales* qui apparaissent naturellement dans les corpus de textes médicaux destinés au grand public. *La reformulation* représente le processus de réécriture qui a le rôle d'expliquer, simplifier ou pointer une phrase ou un syntagme. Nous travaillons sur la *reformulation paraphrastique* qui conserve le sens et va vers une équivalence sémantique (Fuchs, 2020 ; Pennec, 2020 ; Vassiliadou, 2020), mais également sur la *reformulation non paraphrastique* qui exprime un changement de perspective énonciative (Rossari, 1990 ; Fuchs, 1994). Nous suivons le point de vue de Vassiliadou (2020) qui propose une définition plus stricte de la *reformulation* pour exclure les structures beaucoup trop éloignées du sens de base du *formulé* et du *reformulé*, comme dans le cas de la répétition ou de la reformulation corrective (qui revient sur le dit afin d'apporter une correction). Les *reformulations non paraphrastiques* sont prises en compte dans notre étude uniquement si elles permettent de donner une précision, une explication ou une définition du *terme médical*, qui est dans notre cas *le formulé*.

Dans notre étude nous analysons les reformulations de *termes*, des unités lexicales de spécialité qui représentent des connaissances spécifiques à un domaine du savoir, reconnues et partagées par une communauté de spécialistes (Costa, 2005). Nous cherchons toutes les *reformulations* qui peuvent servir à la vulgarisation des textes médicaux. Nous menons notre travail de recherche sur la *reformulation sous-phrastique* (Bouamor, 2012 ; Fuchs, 1982), qui ne dépasse pas la longueur d'une phrase, de type « **myotonie**, *c'est-à-dire* **une sensation de raideur musculaire** ». Nous prenons en compte les spécificités du **texte médical** et des **termes**. L'objectif de la reformulation est la simplification des termes médicaux pour présenter dans un langage plus accessible les informations médicales au grand public (Grabar et Hamon, 2015, 2016 ; Cardon et Grabar, 2018 ; Koptient *et al.*, 2019 ; Cardon, 2021 ; Buhnila, 2022b), pour faciliter la communication avec les patients (Pecout *et al.*, 2019 ; Koptient et Grabar, 2020). Les reformulations peuvent avoir différentes *fonctions* dans le texte, par rapport au message destiné au public cible : définir un terme médical, donner une paraphrase plus facile à comprendre, expliquer une procédure médicale, etc. (Eshkol-Taravella et Grabar, 2017). La reformulation peut contenir des concepts plus spécifiques (*hyponymes*) ou plus génériques que le terme médical (*hyperonymes*), des *synonymes* ou *méronymes* pour simplifier le terme médical reformulé.

Notre travail est mené sur des **corpus bilingues comparables** du domaine de la médecine, en **français et roumain**. Notre objectif est de créer semi-automatiquement une base de termes médicaux accompagnés de leur reformulations pour construire un **système de simplification automatique de termes**. Nous présentons nos corpus de travail, suivis par notre méthodologie d'annotation automatique et manuelle, notre analyse quantitative et qualitative et nos résultats.

Corpus et méthodologie

Corpus

Nous faisons l'hypothèse que les textes grand public contiennent un nombre plus grand de reformulations dans un langage simplifié, afin de rendre ces textes compréhensibles pour les non-spécialistes. Pour la langue française nous travaillons sur le corpus **CLEAR Cochrane** (Grabar et Cardon, 2018). Ce corpus écrit comparable est constitué de textes scientifiques du domaine médical destinés aux experts et de textes simplifiés qui traitent des sujets de la médecine destinés au grand public. Pour nos expériences nous analysons le *corpus pour le grand public* (à partir de maintenant, **CLEAR GP**), qui a une taille de **1 515 051 tokens**. Le travail de nettoyage, anonymisation et prétraitement du corpus a déjà été réalisé. Pour la langue roumaine, nous constituons un corpus de vulgarisation médicale en nous inspirant du corpus **GrandMed-RO** (Buhnila, 2018), corpus roumain des textes scientifiques et de vulgarisation, avec une taille de 42 140 tokens. Nous agrandissons ce corpus à travers l'outil de génération des corpus Sketch Engine (Kilgarriff *et al.*, 2014). Nous avons choisi huit sites différents de vulgarisation médicale[26] qui contiennent un grand nombre de textes et qui permettent l'extraction automatique de données. Le corpus **GrandMed-RO** agrandi avec Sketch Engine est composé de 7 472 articles et a une taille totale de **6 440 951 tokens**.

table 1. :

Méthodologie

Une fois les deux corpus nettoyés et alignés une phrase par ligne, nous avons procédé à l'identification automatique des termes médicaux avec des scripts en langage Perl. Nous avons utilisé pour le français l'ontologie médicale **SNOMED 3.5-VF** (Côté, 1996) qui contient 150 906 concepts médicaux avec l'outil d'annotation automatique **SIFR-BioPortal** (Tchechmedjiev *et al.*, 2018). Pourtant, de tels outils et ressources ne sont pas encore disponibles pour le roumain. Notre méthode consiste dans la construction d'une liste de 14 133 termes médicaux extraits par nous à partir du travail d'annotation d'entités nommées du projet **MoNERo** (Mitrofan *et al.*, 2019). L'annotation des termes médicaux a été validée par des professionnels du domaine médical. Les termes médicaux que nous cherchons automatiquement avec des scripts proviennent des **terminologies** ou **listes des termes** présentées, ils ne sont pas lemmatisés ni annotés morpho-syntaxiquement. Cette méthode permet d'identifier à la fois les termes au pluriel et au singulier pour le français. Pour le

[26]Sites de vulgarisation médicale en roumain : https://sfaturimedicale.ro/ ; https://www.sfatulmedicului.ro/ ; https://www.doctorulzilei.ro/ ; https://www.romedic.ro/ ; https://www.csid.ro/boli-afectiuni/ ; https://www.csid.ro/sanatate/; https://www.cdt-babes.ro/ ; https://www.reginamaria.ro/medici.

roumain, la liste des termes a été extraite à partir de termes annotés manuellement dans le contexte, ce qui justifie la présence de variations en nombre, articles définis et cas.[27]

L'annotation automatique des termes médicaux est suivie par la recherche de **marqueurs de reformulation** également avec des scripts. Notre hypothèse de travail est que les marqueurs lexicaux, grammaticaux, voire orthographiques, permettent d'identifier des reformulations médicales de façon automatique (Buhnila, 2022a). Pour cela, nous constituons une liste de **marqueurs de reformulations** basés sur le verbe « dire », comme « c'est-à-dire », « ça veut dire », « pour dire autrement », « autrement dit » (Vassiliadou, 2013, 2016 ; Steuckardt, 2018 ; Magri, 2018), et autres, de type « signifie », « est ce qu'on appelle », « aussi appelé », « doit être compris comme », « au sens de » (Rey-Debove, 1978). Nous rajoutons nos observations sur les corpus et nous créons une liste équivalente de marqueurs pour la langue roumaine. Une fois les marqueurs identifiés automatiquement avec des scripts, nous analysons les phrases qui contiennent à la fois des termes médicaux et des marqueurs de reformulations.

Une équipe de deux annotateurs a annoté manuellement les reformulations en suivant le **guide d'annotation** que nous avons constitué. Notre guide d'annotation contient des exemples prototypes pour chaque type de donnée à annoter : les termes médicaux (simples ou polylexicaux), les différentes catégories de marqueurs identifiés, le statut de la phrase (si elle contient une ou plusieurs reformulations correctes ou aucune reformulation correcte) et les types de reformulations sous-phrastiques. Pour identifier les **reformulations correctes**, nous avons évalué manuellement les liens entre termes et marqueurs identifiés automatiquement dans les phrases extraites. Parfois, le terme identifié n'est pas celui reformulé, ou le marqueur n'introduit pas une reformulation médicale.

Pour aller plus loin dans l'analyse des reformulations médicales et leur rôle dans la vulgarisation scientifique, nous avons analysé également **les relations lexicales** établies entre la reformulation et le terme (de type *hypéronymie, hyponymie, synonymie, méronymie*) (Condamines, 2018 ; Ramadier, 2016 ; Săpoiu, 2013) et **les fonctions sémantico-pragmatiques** qui indiquent les raisons pour lesquelles les reformulations sont employées dans les textes médicaux. Nous avons considéré que les fonctions sémantico-pragmatiques suivantes sont adaptées aux textes médicaux écrits : donner la *définition* d'un terme, *paraphraser* avec des mots plus simples un terme technique, donner des exemples pour illustrer un terme (*exemplification*), nommer un terme à travers un autre terme (*dénomination*) ou donner une *explication* détaillée (Eshkol-Taravella et Grabar, 2017 ; Buhnila, 2022a). L'évaluation des annotations est réalisée à l'aide d'un accord inter-annotateur de type **Kappa** (Cohen, 1960) et avec des mesures statistiques de précision.

Résultats

Pour le corpus CLEAR GP nous avons 125 696 termes médicaux annotés (73% des phrases ont des termes médicaux) dont 120 450 doublons, ce qui nous donne **5 100** termes uniques après nettoyage des mots erronément identifiés par l'annotateur comme des termes médicaux (comme « après », « en plus de », « suivant », « compatible avec »).

[27] Le roumain a gardé les marques morphologiques enclitiques des cas. Les cas sont hérités du latin (nominatif, accusatif, datif, génitif, vocatif). L'article défini est attaché sous forme de suffixe, par exemple *afecțiuni* (affections) -> *afecțiunile* (les affections).

Mesures statistiques	CLEAR GP				GrandMed-RO	
	Annot 1	**%**	**Annot 2**	**%**	**Annot 1**	**%**
Précision - ***oui***	1462	36,77%	1600	40,25%	2 370	64,70%
Précision - ***non***	2513	63,22%	2375	59,74%	1 293	35,29%
N° ref multiples	347	+8,72%	645	+16,22%	657	+14,12%
N° total ref	**3975**	**100%**	**3975**	**100%**	**3663**	**100%**
Score Kappa	**0,71**					

Table 1. Données et mesures statistiques sur l'annotation des corpus

À l'aide de notre méthode de pré-annotation automatique nous avons identifié en français **3 975** phrases dont **38,51%** (moyenne des deux annotateurs) sont des reformulations médicales correctes. Les annotations ont été réalisées par deux annotateurs francophones non-spécialistes du domaine de la médecine. L'accord inter-annotateur Kappa est de **0,71** qui est un *accord modéré* (McHugh, 2012). Pour le roumain, nous avons annoté les deux premiers sous-corpus, « sfaturi medicale » (*avis médicaux*) et « sfatul medicului » (*l'avis du médecin*), qui ont un taille de **1 960 152** tokens. Nous avons annoté **3 663 phrases** du corpus **GrandMed-Ro**, dont **2 370 phrases** (**64,70%**) contiennent des **reformulations correctes** en roumain.

L'annotation manuelle des relations lexicales et fonctions sémantico-pragmatiques est réalisée sur les phrases dont les discrépances entre les deux annotations ont été harmonisées pour obtenir un total de **1 767** phrases pour le corpus **CLEAR GP** et sur les **2 370** phrases pour **GrandMed-RO**. Nous observons que **49,46%**, respectivement **37%** de reformulations sont annotées avec *hypéronymie-définition,* **26,37%** et **29,32%** sont annotées avec la relation *hyponymie* et avec la fonction *d'exemplification* (donner des sous-catégories d'éléments qui aident à expliquer le terme). L'annotation *synonymie-paraphrase / dénomination* est présente à **16,69%** et **17,67%,** tandis que la paire *méronymie-explication* est peu présente. Pourtant, une nouvelle association, de type *méronymie-définition*, concerne **7,17%** de reformulations dans le corpus roumain.

Deux annotateurs natifs ont analysé les relations lexicales et fonctions sémantico-pragmatiques de **1 000 reformulations correctes** en roumain. Les annotateurs ont été d'accord pour attribuer les mêmes **relations** à **60,80%** de reformulations et les mêmes **fonctions**, séparément de relations, à **60,10%** de reformulations. Concernant les paires de *relations-fonctions*, ils ont attribué les mêmes pour **48,70%** de reformulations.

Relation-Fonction	CLEAR GP	GrandMed-RO
	%	%
hypéronymie-définition	**49,46%**	**37%**
hyponymie-exemplification	26,37%	29,32%
synonymie-paraphrase / dénomination	16,69%	17,67%
méronymie-explication	4,01%	3,37%
méronymie-définition	0%	7,17%

Table 2. Mesures statistiques sur l'annotation avec des relations et des fonctions

Conclusion

Notre méthode prouve que les *textes médicaux pour le grand public* contiennent un grand nombre de *reformulations médicales diverses*, que nous avons identifiées automatiquement avec une précision de **38,51%** pour le corpus français et de **64,70%** pour le corpus roumain. L'amélioration pour ce dernier est due à l'affinement de la *liste de marqueurs de reformulation* en roumain. Notre analyse lexicale et sémantico-pragmatique prouve que les reformulations de type *hypéronymie-définition* sont les plus fréquentes (**49,46%** et **37%**) dans les corpus médicaux écrits de vulgarisation médicale.

Références bibliographiques

Bouamor, H. (2012). Étude de la paraphrase sous-phrastique en traitement automatique des langues. *Orsay : Université Paris Sud - Paris XI.* https://tel.archives-ouvertes.fr/tel-00717702.

Buhnila, I. (2018). Simplification lexicale entre les textes scientifiques et les textes de vulgarisation du domaine de la médecine. *Mémoire de Master, Université de Strasbourg*. F-67000 Strasbourg, France.

Buhnila, I. (2022a). Le Rôle Des Marqueurs et Indicateurs Dans l'analyse Lexicale et Sémantico-Pragmatique de Reformulations Médicales. *8e Congrès Mondial de Linguistique Française (CMLF)*, 4-8 juillet 2022, Orléans, *SHS Web of Conferences* 138: 10005. https://doi.org/10.1051/shsconf/202213810005.

Buhnila, I. (2022b). Identifying Medical Paraphrases in Scientific versus Popularization Texts in French for Laypeople Understanding. In *Proceedings of the Third Workshop on Scholarly Document Processing*, *Association for Computational Linguistics*. Gyeongju, Republic of Korea, 69–79.

Cardon, R. (2021). Simplification automatique de textes techniques et spécialisés. Informatique et langage [cs.CL]. *Université de Lille*. Français. ⟨NNT : 2021LILUH007⟩. ⟨tel-03343769v2⟩.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas*., 20, 27-46.

Condamines, A. (2018). Nouvelles perspectives pour la terminologie textuelle. J. Altmanova; M. Centrella; K.E. Russo. *Terminology and Discourse, Peter Lang*, 1-13. 978-3-0343-2415-1. ff10.3726/978-3-0343-2414-4ff. ffhalshs-01899150f.

Costa, R. (2005). Texte, terme et contexte. In *Actes des septièmes Journées scientifiques du réseau de chercheurs Lexicologie Terminologie Traduction*. Bruxelles, Belgique, 79-88.

Côté, R. (1996). Répertoire d'anatomopathologie de la SNOMED internationale, v3.4. *Université de Sherbrooke*, Sherbrooke, Québec.

Eshkol-Taravella, I. et Grabar, N. (2017). Taxinomie dans les reformulations du point de vue de la linguistique de corpus. *Syntaxe et Sémantique, vol. 18, no. 1*, 149-184.

Fuchs, C. (1982). La Paraphrase. *PUF*. Paris, 184 pages.

Fuchs, C. (1994). Paraphrase et énonciation. *Éditions OPHRYS*, 185 pages.

Fuchs, C. (2020). Paraphrase et reformulation : un chassé-croisé entre deux notions. *Olga Inkova (dir). Autour de la reformulation, 36, Droz*, 41-55, Coll. Recherches et Rencontres, 978-2-600-06051-6.

Grabar, N. et Cardon, R. (2018). CLEAR - Simple Corpus for Medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA),* Tilburg, the Netherlands: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-7002, *3–9*.

Grabar, N. et Hamon, T. (2015). Extraction automatique de paraphrases grand public pour les termes médicaux. In *22ème Traitement Automatique des Langues Naturelles*, 14. Caen, France.

Grabar, N. et Hamon, T. (2016). Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux. *Traitement Automatique des Langues, Varia, 57 (1)*: 85-109.

Kilgarriff, A., Baisa V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. et Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography vol. 1*, 7-36.

Koptient, A., Cardon, R. et Grabar, N. (2019). Simplification-induced transformations: typology and some characteristics. In *Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics*. Florence, Italy. https://doi.org/10.18653/v1/W19-5033. 309–318.

Koptient, A. et Grabar, N. (2020). Rated Lexicon for the Simplification of Medical Texts. *The Fifth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing HEALTHINFO 2020*, Oct 2020, Porto, Portugal. ⟨hal-03095275⟩.

Magri, V. (2018). Marqueurs de reformulation : exploration outillée et contrastive dans deux corpus narratifs. *Langages N° 212 (4):* 35-50.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica, 22(3)*, 276–282.

Mitrofan, M., Barbu Mititelu, V. et Mitrofan, G. (2019). MoNERo: A Biomedical Gold Standard Corpus for the Romanian Language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, Association for Computational Linguistics, Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5008, 71-79.

Pecout, A, Tran, T. M. et Grabar, N. (2019). Améliorer la diffusion de l'information sur la maladie d'Alzheimer : étude pilote sur la simplification de textes médicaux. *Ela. Etudes de linguistique appliquée N° 195 (3)*: 325 41.

Pennec, B. (2020). Les reformulations : des formes méta-énonciatives par excellence. Spécificités et introducteurs. *Olga Inkova (dir). Autour de la reformulation, 36, Droz*, 57-75, Coll. Recherches et Rencontres.

Ramadier, L. (2016). Indexation et apprentissage de termes et de relations à partir de comptes rendus de radiologie. *Informatique. Université Montpellier*, Français. ⟨NNT: 2016MONTT298⟩. ⟨tel-01479769v2⟩.

Rossari, C. (1990). Projet pour une typologie des opérations de reformulation. *Cahiers de linguistique française 11,* 345-359.

Săpoiu, C. (2013). Hiponimia în terminologia medicală. Modalităţi de abordare în semantică şi lexicografie. *Editura Trend*, Piteşti,199 pages.

Steuckardt, A. (2018). Les marqueurs de paraphrase formés sur dire : exploration outillée. *Langages N° 212 (4)*, 17-34.

Tchechmedjiev, A., Abdaoui, A., Emonet, V., Zevio, S. et Jonquet, C. (2018). SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes. *BMC bioinformatics, 19(1)*, 405.

Vassiliadou, H. (2013). C'est-à-dire (que) : embrayeur d'énonciation. Semen. *Revue de sémio-linguistique des textes et discours, no 36 (octobre).* 1-14. http://journals.openedition.org/semen/9684. https://doi.org/10.4000/semen.9684.

Vassiliadou, H. (2016). Mouvements de réflexion sur le dire et le dit : c'est-à-dire, autrement dit, ça veut dire. *Histoires de dire. Petit glossaire des marqueurs formés sur le verbe dire, L. Rouanne & J.-C. Anscombre (éds)*, Bern/Berlin/Bruxelles/New York/Oxford/Wien, Peter Lang, 339-364.

Vassiliadou, H. (2020). Peut-on aborder la notion de "reformulation" autrement que par la typologie des marqueurs ? Pour une analyse sémasiologique et onomasiologique. *Olga Inkova (dir). Autour de la Reformulation, 36, Droz*, 77-94, 978-2-600-06051-6.

Appréhender la production du langage oral en école maternelle en croisant les focales

Laurence Buson [1], Solange Rossato [2] et Isabelle Rousset [1]

[1] Laboratoire LIDILEM, Université Grenoble Alpes

[2] Laboratoire LIG, Université Grenoble Alpes

Introduction

Le rôle prépondérant du langage oral en école maternelle est largement pris en compte dans les programmes de l'Education Nationale[28]. En effet, de nombreuses études ont mis en évidence que les habiletés langagières des jeunes enfants constituent des prédicteurs majeurs de la réussite scolaire ultérieure (Florin, 1991; Lahire, 1993). Une part importante de ces disparités langagières est corrélée au milieu social d'origine des enfants (Bara et al., 2008; Daussin et al., 2011; Écalle et al., 2020; Hart & Risley, 2003; Hoff, 2003; Huttenlocher et al., 2010; Nardy et al., 2013; OCDE, 2016), illustrant ainsi le rôle crucial du langage dans la reproduction des inégalités sociales.

Ces disparités langagières sont largement attestées en école maternelle, notamment en terme de quantité de parole avec la notion de « petits parleurs » (Florin, 1991; Lentin, 1972). D'autres mesures des compétences langagières s'appuient sur l'allongement des longueurs d'énoncés, la taille du lexique et sa diversité, l'utilisation de mots grammaticaux ou encore la diversité des structures syntaxiques. Ces informations sont parfois recueillies auprès des parents ou des adultes s'occupant des enfants (Bassano et al., 2020; Rousset et al., 2019), d'autres proviennent d'enregistrements de dyades adulte-enfant, par exemple autour d'un jeu (Parisse & Le Normand, 2007), tandis que d'autres études se basent sur des ateliers langage en école (Péroz, 2016).

Les enseignants en maternelle, confrontés à la grande diversité des niveaux langagiers des enfants, manquent « d'outils et de moyens pour observer, programmer et évaluer les productions langagières » (Canut et al., 2013, 44). Un des premiers enjeux est de trouver une situation qui favorise la prise de parole des enfants. Rousset et al. (2019) étudient ainsi un dispositif d'ateliers langage en petits groupes homogènes autour d'un album de littérature jeunesse. Par ailleurs, plusieurs études (Boisseau, 2005; Canut, 2009 ; Lentin, 1972) soulignent l'importance du travail de la morpho-syntaxe. C'est sur ces bases que nous avons mis en œuvre une recherche-action à destination des terrains scolaires dans le cadre du projet PARM[29] financé par Pégase[30] (PIA3). Des albums ont été spécialement conçus en s'appuyant

[28] Voir par exemple la version des programmes entrée en vigueur en 2021. Consultable ici : https://cache.media.education.gouv.fr/file/25/86/5/ensel550_annexe_1413865.pdf

[29] https://www.polepilote-pegase.fr/recherche/rd-collaborative/projet-parm-parler-raconter-en-maternelle/

[30] Opération soutenue par le pôle Pégase et l'Etat, dans le cadre de l'action « Territoires d'innovation pédagogique » du programme d'investissement d'avenir, opéré par la Caisse des Dépots.

sur une progression morphosyntaxique et une complexification progressive des structures syntaxiques proposées aux élèves, tout en élargissant leur palette stylistique au fil de la progression (Buson, 2010; Buson & Nardy, 2014, 2020).

Dans le cadre de ce projet, la question de l'évaluation des compétences langagières des enfants est posée au travers de différentes focales qui croisent des informations déclaratives des enseignants, une situation de dyade adulte-enfant qui consiste à raconter une histoire à partir d'images séquentielles, et différents ateliers réalisés en classe autour d'albums et de jeux spécifiquement conçus dans le cadre de cette recherche-action.

Cette communication propose de présenter le dispositif, le recueil de données, ainsi que les différents types d'analyses et de mesures qui peuvent être faites pour chaque situation, pour ensuite poser la question de leur mise en synergie. Nous proposerons également les premiers résultats obtenus sur la base d'un sous-corpus de données.

Dispositif pour les ateliers langage

Le dispositif mis en œuvre dans les classes consiste en un travail spécifique centré sur la production orale, réalisé autour d'albums et de jeux, dont l'objectif est d'adapter les structures langagières ciblées au niveau de développement langagier des enfants, rassemblés en petits groupes homogènes. Le contenu des albums (longueur des phrases, temps des verbes, type d'interrogatives, énoncés enrichis et complexes) est pensé en fonction des niveaux langagiers des enfants, selon une progression cible qui va de 7 (pour l'album le plus simple) à 28 (pour l'album le plus complexe) selon une grille de calcul de score morphosyntaxique (grille « MS », présentée en Annexe 1) élaborée dans le cadre de ce projet pour permettre une évaluation plus précise des compétences syntaxiques des élèves[31]. Les textes des albums ont été écrits par l'une des chercheuses du projet, illustrés par un illustrateur et deux illustratrices[32] et édités par UGA Edition. Les contenus langagiers des albums ainsi que les jeux et leurs modalités de différenciation ont été travaillés avec des enseignantes et des conseillères pédagogiques[33]. Les enseignantes volontaires cette année[34] pour tester le dispositif ont mis en œuvre les 6 séances prévues de la séquence pédagogique (cf. séquence en Annexe 2) lors de la période 3 (janv-fév 23). Elles ont pu expérimenter quatre albums/kits de jeux sur les huit prévus pour la version finale.

Question de recherche

Dans cette communication, notre questionnement de recherche concerne la manière dont différentes modalités d'évaluation des compétences langagières des enfants peuvent se "répondre", se compléter et s'articuler. En effet, la validation de ces outils d'évaluation dépend

[31] Cette grille sera amenée à évoluer en fonction des analyses des productions des enfants, pour en proposer une version finale diffusable en fin de projet.

[32] Georges Crisci, Clothilde Keraudran et Céline Vinante.

[33] Anne-Cécile Despinasse et Frédérique Bouvier, conseillères pédagogiques à Grenoble et Voiron, ainsi que Valérie Battistini, Phyllis Graff, et Laura Terpant, enseignantes de maternelle.

[34] Nous utiliserons le féminin pour les désigner dans ce résumé, vu qu'il s'agit de 19 femmes et de 1 homme : Chloé Portz, Marion Bourdin, Cécile Richard Didry, Romy Brun, Bénédicte Gautheron, Cyliane De Waele, Claire Heintz, Aude Bard, Sylvie Rizzi, Marie Philippe, Camille Piras, Laetitia Colonna, Caroline Riassetto, Mariette Tholence, Juliette Pauly, Sophie Cuaz, Cristel d'Oria, Camille Daumas, Amandine Lorne et Guillaume de Petigny. Nous remercions également Nathalie Penin, Inspectrice de l'Education Nationale et Marie-Line Cauquil, conseillère pédagogique, pour leur accompagnement dans la mise en œuvre de l'expérimentation.

de notre capacité à identifier les corrélations entre les étapes de développement indiquées par les enseignants et les scores morphosyntaxiques obtenus à partir du traitement automatique des transcriptions des ateliers en classe.

Corpus

Le recueil de données dans le cadre du projet PARM concerne 5 écoles d'indice de position sociale inférieur à la moyenne (IPS entre 84 et 102, tandis que la moyenne dans l'Académie est à 109), soit 15 classes et près de 300 élèves. Le corpus comprend :

i) les grilles de positionnement individuelles[35] remplies par les enseignantes avant (temps 1 : novembre) et après (temps 2 : avril) le travail autour des albums, cette grille ayant déjà été testée lors de précédentes expérimentations (Rousset et al. 2019) ;

ii) des enregistrements vidéos, réalisés sur le temps 1, de dyades expérimentateur/enfant selon un protocole standardisé (tâche de narration à partir de quatre images avec une amorce identique pour tous) ; ces vidéos permettent d'attribuer un score morphosyntaxique préalable à l'expérimentation pour chaque enfant ;

iii) des enregistrements vidéos d'ateliers langage en petits groupes homogènes, réalisés en classe par les enseignantes elles-mêmes à l'aide de caméras 360°. Ils concernent des séances de narration collaborative à l'aide de marottes reprenant les personnages des albums, correspondant aux ateliers 2 et 5 (cf. Annexe 2).

La présente communication se focalisera sur deux types de données parmi celles listées précédemment : les évaluations fournies par les enseignantes via les grilles de positionnement, et les traitements automatiques issus des transcriptions et annotations manuelles des vidéos[36] d'ateliers. A ce stade du traitement des données, le corpus disponible pour l'analyse concerne à peu près la moitié de notre corpus final, soit 139 enfants répartis dans 8 classes (54 TPS-PS, 29 MS, et 56 GS pour des âges allant de 2;8 à 5;9 au temps 1 de l'étude).

Méthodologie de l'analyse

Les transcriptions et annotations utilisent le logiciel ELAN (2022) et une grille d'annotation élaborée par l'équipe de recherche (cf. Annexe 3) permettant de détailler les productions des élèves notamment au niveau de la construction des énoncés. Une fois ce *template* rempli manuellement par les transcripteurs dans ELAN, un script PRAAT permet de renseigner automatiquement la grille MS (cf. Annexe 1). Pour chaque enfant, nous croiserons donc différents indicateurs de niveau langagier[37] : l'étape de développement entre 0 et 4 issue de la grille de positionnement remplie par les enseignants aux temps 1 et 2, et les scores de compétences morphosyntaxiques compris entre 0 et 28 respectivement pour l'atelier 2 et l'atelier 5.

Notre communication décrira la démarche, les méthodologies de recueil et d'analyses, et présentera quelques premiers résultats issus du traitement du sous-corpus de 7 classes. Nous proposerons ainsi les premières pistes d'une exploitation conjointe de différents types de

[35] https://lidilem.univ-grenoble-alpes.fr/sites/lidilem/files/Mediatheque/Ressources/grille_positionnement_vf.pdf

[36] Les transcriptions et codages sont réalisés par une équipe de stagiaires et vacataires, préalablement formé·es au logiciel ELAN et à l'utilisation de la grille d'annotation : Ivonne Saldivar-Rodriguez, Noah Jourdan, Lucille Mariani, Claire Praly et Pierre Romangin.

[37] Cette grille MS a également été utilisée par les enseignantes pour évaluer les productions de leurs élèves lors des ateliers jeux, mais ces données ne seront pas prises en compte dans la présente communication.

données, avec en filigrane un questionnement méthodologique sur la complémentarité possible de données langagières de différentes natures dans le cadre de corpus au traitement complexe tels que les corpus de données orales enfantines.

Références bibliographiques

Bara, F., Gentaz, É., & Colé, P. (2008). Littératie précoce et apprentissage de la lecture : Comparaison entre des enfants à risque, scolarisés en France dans des réseaux d'éducation prioritaire, et des enfants de classes régulières. *Revue des sciences de l'éducation*, 34(1), 27-45.

Bassano, D., Labrell, F., & Bonnet, P. (2020). *Évaluer les débuts du langage avec le DLPF. Lexique, grammaire et pragmatique chez le jeune enfant* (halshs-02436994). https://halshs.archives-ouvertes.fr/halshs-02436994/document

Boisseau, P. (2005). *Enseigner la langue orale en maternelle.* Paris, Retz.

Buson, L. (2010). La didactique du FLM, du FLE et du plurilinguisme au service de l'éveil aux styles à l'école : Des pistes pour la formation des enseignants. Actes du Congrès Mondial de Linguistique Française (CMLF). https://doi.org/10.1051/cmlf/2010117

Buson, L., & Nardy, A. (2014). Repenser l'enseignement/apprentissage des registres de langue en français : Quelles articulations entre acquisition, sociolinguistique et didactique ? Travaux de didactique du français langue étrangère, 67-68, 147-163.

Buson, L., & Nardy, A. (2020). Le(s) français scolaire(s) à l'école maternelle. Pratiques langagières d'enseignant.e.s et d'élèves. Le français aujourd'hui, 208, 93-103.

Canut, E. (2009). Apprendre à parler pour ensuite apprendre à lire et à écrire. In *Congrès FNAME, Le langage. Objet d'apprentissage, outil de pensée. Quels obstacles? Quels leviers?*.

Canut, E., Espinosa, N., & Vertalier, M. (2013). Corpus et prise de conscience des processus interactionnels d'apprentissage du langage pour repenser les pratiques enseignantes en maternelle. *Linx. Revue des linguistes de l'université Paris X Nanterre* (68-69), 69-93.

Daussin, J.-M., Keskpaik, S., & Rocher, T. (2011). L'évolution du nombre d'élèves en difficulté face à l'écrit depuis une dizaine d'années. *France, Portrait social*, 137-152.

ELAN (Version 6.4) [Computer software]. (2022). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Écalle, J., Labat, H., Thierry, X., & Magnan, A. (2020). Évaluation des compétences en littératie chez les enfants français de 4-5 ans. *Sante Publique,* Vol. 32(1), 9-17.

Florin, A. (1991). *Pratiques du langage à l'école maternelle et prédiction de la réussite scolaire* (Presses Universitaires de France).

Lahire, B. (1993). *Culture écrite et inégalités scolaires. Sociologie de l'"échec scolaire" à l'école primaire.* Presses Universitaires de Lyon.

Lentin, L. (1972). *Comment apprendre à parler à l'enfant de moins de six ans. Où ? Quand ? Comment ?* ESF.

Hart, B., & Risley, T. R. (2003). The early catastrophe : The 30 million word gap by age 3. *American Educator*, 27(1), 4-9.

Hoff, E. (2003). The specificity of environmental influence : Socioeconomic status affects early vocabulary development via maternal speech. *Child Development,* 74(5), 1368-1378.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology,* 61(4), 343-365.

Nardy, A., Chevrot, J.-P., & Barbu, S. (2013). The acquisition of sociolinguistic variation : Looking back and thinking ahead. *Linguistics*, 51(2), 255-284. https://doi.org/10.1515/ling-2013-0011

OCDE. (2016). *Résultats du PISA 2015* (Volume I). https://doi.org/10.1787/9789264267534-fr

Parisse, C., & Le Normand, M.-T. (2007). Une méthode pour évaluer la production du langage spontané chez l'enfant de 2 à 4 ans. *Glossa,* UNADREO-Union Nationale pour le Développement de la Recherche en Orthophonie (97), 10-30.

Péroz, P. (2016). Apprentissage du langage oral à l'école maternelle. *Pratiques* [En ligne], 169-170. http://pratiques.revues.org/3100 ; DOI : 10.4000/pratiques.3100

Rousset, I., Rossato, S., Lequette, C., & Latapie, E. (2019). Exploration des compétences langagières des enfants d'écoles maternelles en zone d'éducation prioritaire. *10èmes Journées Internationale de la Linguistique de Corpus*, Grenoble. https://hal.archives-ouvertes.fr/hal-02457055

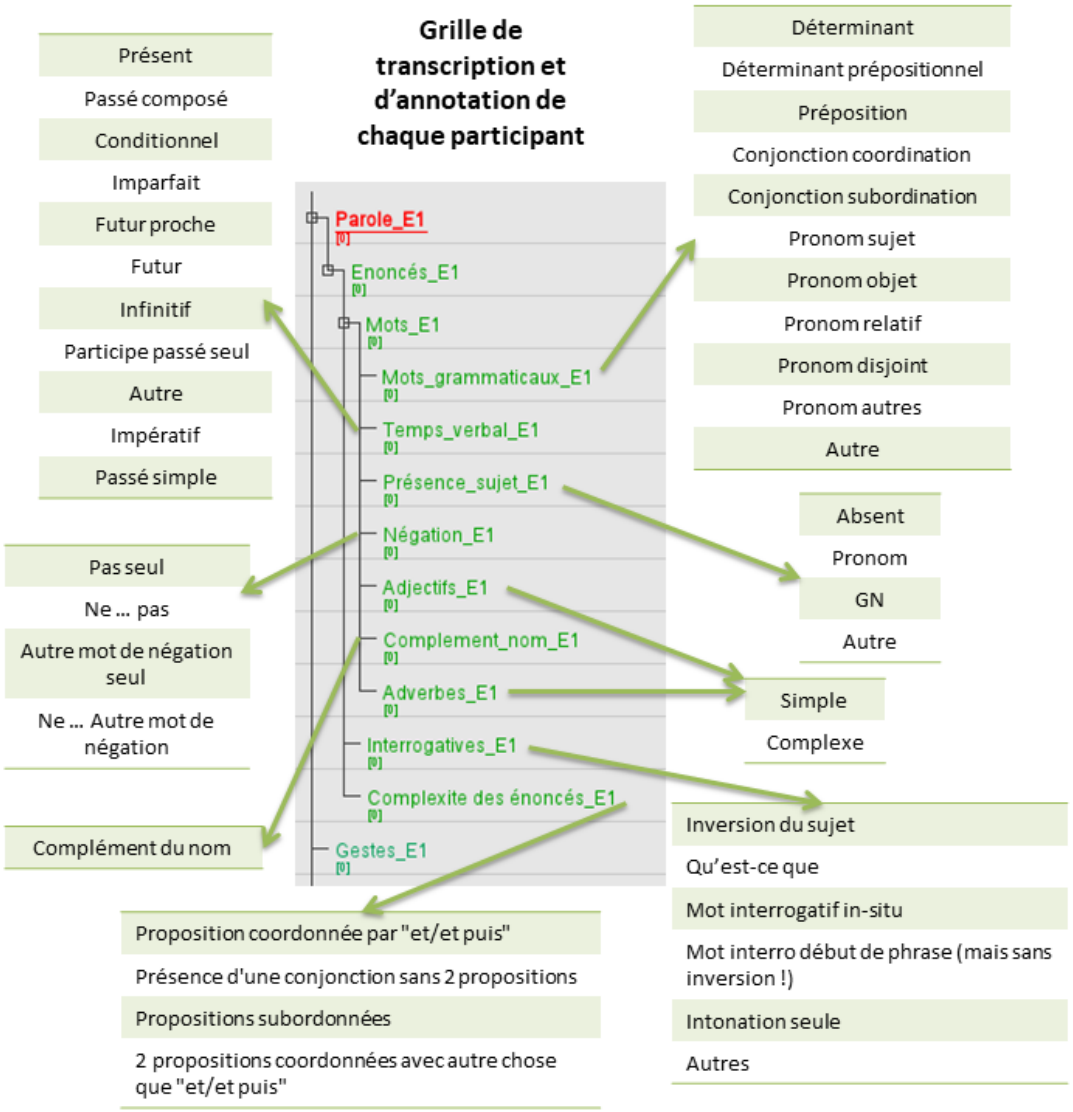
Annexe 1 : Grille morphosyntaxique (« grille score MS »)

	0 point	1 point	2 points	3 points	4 points	Scores T1	Scores T2
Longueur des énoncés	silence ou non compréhensible	mots isolés	combine 2 mots	combine 3 - 4 mots	combine 5 mots ou +		
Enoncés enrichis (présence d'adjectifs, adverbes, compléments du nom)	pas d'énoncé enrichi	1 seule catégorie simple parmi adjectifs simples (=*petit/grand*) ou adverbe simples (=*oui, non, très, trop, aussi, beaucoup*)	1 seule catégorie parmi adjectifs complexes ou adverbes complexes	2 catégories différentes parmi adjectifs, adverbes, compl. du nom	présence des 3 catégories parmi adjectifs, adverbes, compl. du nom		
Mots grammaticaux	0 ou 1 sorte de mot gram.	2 ou 3 sortes de mots gram.	4 ou 5 sortes de mots gram.	6 sortes de mots gram. et +	6 sortes de mots gram. et + avec au moins un pronom relatif ou une conjonction de subordination		
Enoncés complexes	pas d'indicateur de complexité	présence de deux propositions coordonnées par *et/et puis*	présence de proposition subordonnée sans principale, avec un seul verbe conjugué (ex : *parce que je veux pas*)	présence de deux propositions coordonnées par autre chose que *et/et puis*, ou de subordonnées	présence de propositions coordonnées et de subordonnées		
Verbes	pas de verbes conjugués ou participes passés seuls type "parti"	verbes conjugués sans sujets (hormis impératif)	verbes conjugués au présent de l'indicatif avec sujets	verbes conjugués avec sujets parmi passé composé, futur proche, impératif, conditionnel	verbes conjugués avec sujets, présence d'imparfait, de passé simple ou de futur		
Négation	non observé	mot 2 (*ex : pas/plus/rien…*) + infinitif	verbe + mot 2	sujet + verbe + mot 2	sujet + ne + verbe + mot 2		
Interrogatives	non observé	avec intonation seule ou mot interrogatif seul (*ex : quand ?*)	au moins une question avec mot interrogatif, (*quel qu'il soit et quelle que soit sa place*)	2 sortes d'interrogatives, sans inversion	présence d'une interrogative avec inversion du sujet		
						Tot :	**Tot :**

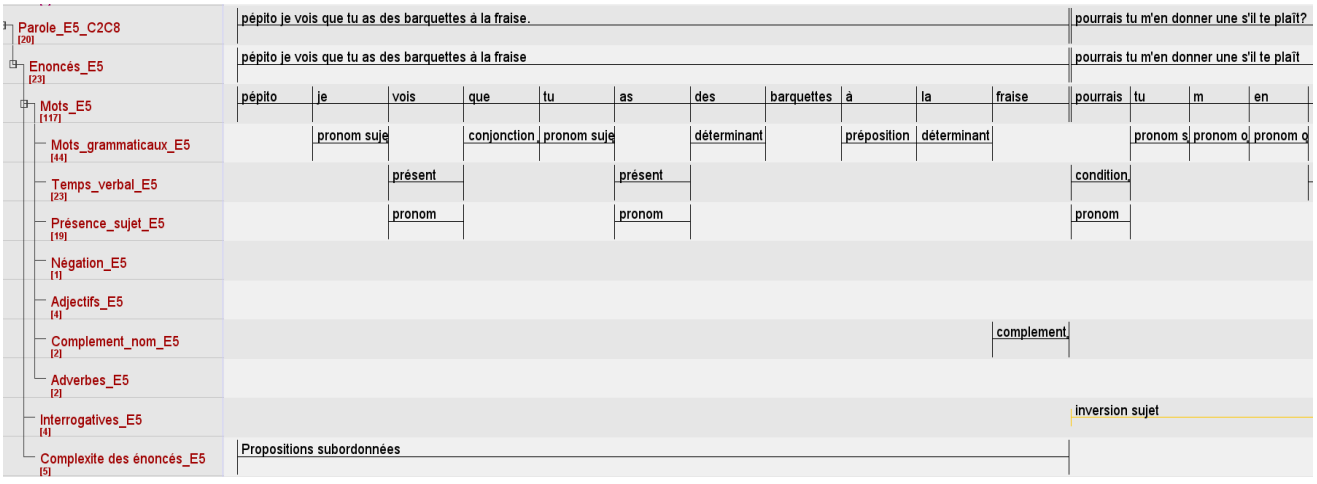
Annexe 2 : Trame de séquence PARM

DÉCOUVERTE	**Atelier 1** : se familiariser avec l'album Objectif : se familiariser avec l'histoire, sa chronologie, le texte, et avec le matériel Rq : possibilité de faire cette étape en collectif ou en ateliers
APPRENTISSAGE	**Atelier 2** : manipuler pour prendre sa place dans la narration Objectif : à l'aide du matériel, commencer à entrer dans la narration collaborative (raconter à plusieurs), prendre sa place dans la trame narrative
APPRENTISSAGE **DIFFÉRENCIÉ**	**Atelier 3_séance jeu décrochée** : travailler les structures cibles en jouant Objectif : travailler de manière différenciée la/les structure(s) cible(s)
APPRENTISSAGE **INSTITUTIONNALISATION**	**Atelier 4** : produire du texte à l'oral (PS), dicter à l'adulte (à partir de la MS) Objectif : produire un « oral écrivable », le plus proche possible du modèle syntaxique fourni par le récit initial ; inventer l'étape manquante de l'histoire pour compléter l'album en fonction des illustrations NB : chaque album comporte une page sans texte.
RÉINVESTISSEMENT	**Atelier 5** : produire une narration collaborative, avec ou sans matériel Objectif : chaque enfant joue un personnage/ une saynète, ou raconte une page du livre ; selon les groupes, les élèves ont ou non recours au matériel. Rq : l'objectif pour les élèves est d'être capable de "publier" cette narration (la rendre publique, la donner à voir à d'autres) : les élèves pourront raconter l'histoire à un autre groupe, à une autre classe, à leurs parents (albums petit format en circulation dans les familles).
RÉINVESTISSEMENT **ÉVALUATION** **DIFFÉRENCIÉS**	**Atelier 6_séance jeu évaluation** : jouer, réinvestir Objectif : reprendre les activités décrochées différenciées de l'atelier 3 pour évaluer l'appropriation des structures cibles par les enfants. La grille "score MS" pourra alors être complétée.

Annexe 3 : Grille d'annotation sous ELAN



Exemple de parole transcrite et annotée



La phraséologie du lexique de l'armement : étude diachronique dans deux corpus romanesques outillés des 19e et 20e siècles

Timothée Celeyron et Julie Sorba
Univ. Grenoble Alpes, LIDILEM
timothee.celeyron@univ-grenoble-alpes.fr , julie.sorba@univ-grenoble-alpes.fr

Introduction

Notre proposition relève des champs de la phraséologie et de la linguistique de corpus outillée. Elle développe ses analyses dans une perspective diachronique. L'intérêt de ce choix réside dans la possibilité d'observer les changements linguistiques à l'œuvre au sein d'un genre textuel (ici le genre romanesque) dans le but de dégager des unités phraséologiques caractéristiques de ce genre (Siepmann 2015, 2016 ; Sorba 2022). Notre objectif précis est d'étudier les collocations dont le pivot est un mot relevant du lexique de l'armement guerrier. L'extraction des données est réalisée à l'aide de l'outil Lexicoscope 2.0 (Kraif 2016, 2019). Elle est effectuée sur deux corpus syntaxiquement arborés : phraseorom19e_fr (romans français du 19e s. ; voir Sorba 2022 : 126-130) et phraseorom_fr_fr (romans français des 20e et 21e s. ; voir Diwersy et al. 2021). Cette étude pilote s'inscrit dans une plus vaste recherche qui procède à des analyses dans une diachronie longue (du latin classique du 1er siècle av. n.e. jusqu'au français contemporain) au sein de textes narratifs.

Cadre théorique

Dans la lignée de Siepmann (2016), nous considérons que « la surreprésentation statistiquement significative de certains phraséologismes pourrait jouer un rôle non négligeable dans la construction littéraire du texte » (p. 22). Cette conception repose sur le fait de considérer la cooccurrence comme un aspect central de la textualité à l'instar de Viprey (2006). Ces travaux sur la nature lexico-grammaticale des textes littéraires sont en plein essor (par ex., Legallois, Charnois & Poibeau 2016) et bénéficient des nombreuses recherches, en linguistique de corpus outillé, effectuées jusqu'alors sur des textes relevant d'autres genres textuels, scientifiques (par ex. Jacques & Tutin 2018) ou professionnels (par ex., Née, Sitri & Véniard 2014).

L'unité phraséologique que nous avons choisie est la collocation, définie comme une association lexicale privilégiée mise en œuvre dans une relation syntaxique, dont la saisie s'opère à l'interface de la sémantique et de la syntaxe (Tutin 2010, p. 13-14). Elle est envisagée comme un « sous-ensemble productif d'expressions lexicalisées binaires, organisées autour d'une structure prédicat-argument. » (Tutin 2013, p. 61). Notre conception de la collocation est à la fois quantitative et qualitative puisqu'elle combine des éléments statistiques avec une analyse fine des contraintes qui pèsent sur les lexies (sur ce point, voir Hausmann & Blumenthal 2006, p. 3).

Corpus, méthodologie et sélection des données

Le premier corpus, phraseorom19e_fr, comporte 67 romans français du 19e siècle (112 190 327 tokens), répartis par décades entre 1830 et 1899. Le second corpus, phraseorom_fr, est constitué de 1621 romans français (103 809 358 tokens), publiés entre 1950 et 2016, partitionnés en 6 sous-corpus selon le sous-genre romanesque représenté (général, fantasy, historique, sentimental, science-fiction, policier). Notre outil de fouille, le Lexicoscope, permet d'extraire un lexicogramme, soit la liste des cooccurrents les plus significatifs, à partir des pivots relevant de notre champ lexical (voir les annexes).

Par « armement guerrier », nous entendons tout objet, ustensile ou projectile dont la conception et la nature est de blesser ou tuer dans le cadre d'un combat guerrier et humain. Cette définition exclut ainsi les armes improvisées ou armes par destination, le matériel de protection ainsi que les armes explosives (*bombe*, *grenade*, etc.) car leur apparition récente n'est pas compatible avec une étude linguistique en diachronie longue. Les modalités de combat restent sensiblement les mêmes au travers des siècles avec des armes de poing et les armes de jet (la distance entre l'assaillant et sa cible reste similaire avec des flèches ou des armes à feu), alors que les armes explosives apportent une dimension nouvelle au combat. Nous pouvons d'ores et déjà établir une distinction au sein de ce lexique entre, d'une part, le lexique générique qui se restreint à *arme*, *lame*, et quelques autres, et de l'autre, le lexique spécifique, que nous pouvons classer selon les outils (*épée*, *dague*, *pistolet*, *fusil*, etc.) et les projectiles (*flèche*, *balle*, *boulet*, etc.).

Résultats

Nous avons choisi d'explorer la combinatoire les deux mots pivots, *épée* et *fusil*, une arme de poing et une arme à feu. Le tableau 1 ci-dessous, réalisé à partir des lexicogrammes fournis en annexes 1 et 2, propose un classement des collocatifs verbaux, nominaux et adjectivaux pour le pivot en opérant une comparaison entre les deux corpus.

Verbes dont le pivot est objet			**Noms modifiés par le pivot**			**Adjectifs qui complètent le pivot**		
Commun	**19e**	**20-21e**	**Commun**	**19e**	**20-21e**	**Commun**	**19e**	**20-21e**
tirer *manier*	*servir* *saisir* *jeter* *suspendre* *croiser* *sauter* *rendre* *porter* *tenir* *mettre* *prendre* *passer* *faire*	*dégainer* *brandir* *rengainer* *armer* *lâcher* *porter* *ramasser* *planter* *forger* *pointer*	*coup* *pommeau* *lame* *pointe* *côté* *garde*	*homme* *bout* *gens*	*poignée* *fil* *plat* *tranchant*	*long*	*nu* *vieux* *grand*	*court* *double* *bâtard* *brisé*

table 2. : Les collocatifs nominaux, verbaux et adjectivaux du pivot *épée* dans les deux corpus.

Nous observons que les collocatifs nominaux sont plus nombreux à être partagés dans les deux corpus que les collocatifs verbaux et adjectivaux qui semblent plus spécifiques à chaque corpus. Ainsi, *tirer* et *manier* sont les deux seuls collocatifs verbaux communs (voir ex.1) alors que 13 et 11 verbes différents apparaissent comme représentatifs de la combinatoire d'*épée* respectivement dans les romans du 19e s. et du 20-21e s. Ce premier résultat est plutôt contre intuitif car on aurait pu penser que les romans du 20-21e s. qui mettent en scène des combats à l'épée, romans historiques (HIST) ou de fantasy (FY) en majorité, se seraient fortement inspirés des romans de la période précédente dans le choix des expressions.

(1) **Tirant son épée**, Pixel frappa de taille, balafrant la demi-figure qui disparut dans un grand cri. (A. Raphaël, *Avant le déluge*, 2011 – FY)

Le tableau 2 ci-dessous propose, sur le même modèle, le classement des collocatifs verbaux, nominaux et adjectivaux du pivot *fusil* (lexicogrammes en annexes 3 et 4).

Verbes dont le pivot est objet			Noms modifiés par le pivot			Adjectifs qui complètent le pivot		
Commun	**19e**	**20-21e**	**Commun**	**19e**	**20-21e**	**Commun**	**19e**	**20-21e**
charger	*jeter* *armer* *faire emporter* *prendre* *avoir* *donner* *tenir*	*armer* *braquer* *épauler* *brandir* *pointer* *tenir* *poser* *porter* *ramasser* *tirer* *décrocher* *lâcher* *menacer* *appuyer* *prendre* *empoigner* *échapper* *?* *coucher*	*canon* *crosse* *chien*	*Coup* *portée* *pierre*	*Balle* *lunette* *cartouche* *autre*	*aucun*	*aucun*	*mitrailleur* *armé* *automatique* *vieux* *bon* *chargé*

table 3. : Les collocatifs nominaux, verbaux et adjectivaux du pivot fusil dans les deux corpus.

Nous observons un phénomène similaire au précédent à savoir que les collocatifs verbaux communs se restreignent à un seul terme (*charger*) alors que 8 apparaissent comme spécifiques aux romans du 19e s. et 18 à ceux du 20-21e s. La combinatoire adjectivale est assez singulière car c'est seulement dans le

corpus du 20e-21e s. que des adjectifs apparaissent comme spécifiques au nom *fusil*. La combinatoire nominale semble plus équilibrée car nous comptons presque le même nombre de collocatifs en commun (3 items : *canon*, *crosse*, *chien ;* voir ex. 2) et spécifiques à chaque corpus (3 ou 4 items).

(2) On se rappelle qu'au moment même où, grâce au bienheureux rayon de soleil, il avait aperçu le canon du fusil, il s'étonnait de la longanimité de Son Éminence à son égard. (A. Dumas, *Les trois Mousquetaires*, 1844)

Ces premières observations nous incitent à poursuivre l'analyse de la combinatoire de ces deux noms *épée* et *fusil* et de son évolution que nous présenterons lors des JLC 2023.

Références bibliographiques

Kraif, O., Diwersy, S. (2012). Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques. TALN 2012, Grenoble, France, 399-406. <hal-01073693>

Kraif, O. (2016). Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. Cahiers de Lexicologie, 108, 91-106. <10.15122/isbn.978-2-406-06281-3.p.0091>

Kraif, O. (2019). Explorer la combinatoire lexico-syntaxique des mots et expressions avec le Lexicoscope. *Langue française*, 203, 67-83. <10.3917/lf.203.0067>

Diwersy, S., Gonon, L., Goossens, V., Kraif, O., Novakova, I, Sorba, J., Vidotto, I. (2021). La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase. *Corpus*, 22, <10.4000/corpus.6101>

Hausmann, F.J., Blumenthal, P. (2006). Présentation : collocations, corpus, dictionnaires. *Langue française*, 150, 3-13. <10.3917/lf.150.0003>

Jacques, M.-P., Tutin, A. (dir.) (2018). *Lexique transversal et formules discursives des sciences humaines*. ISTE Editions.

Legallois, D., Charnois, T., Poibeau, T. (2016). Repérer les clichés dans les romans sentimentaux grâce à la méthode des motifs. *LIDIL*, 53, 95-117. <10.4000/lidil.3950>

Née, É., Sitri, F., Veniard, M. (2014). Pour une approche des routines discursives dans les écrits professionnels. *4e Congrès Mondial de Linguistique Française*, vol. 8, SHS Web of Conférences (p. 2113-2124). EDP Sciences. <10.1051/shsconf/20140801195>

Siepmann, D. (2015). A corpus-based investigation into key words and key patterns in post-war fiction. *Functions of language* 22.3, 362-399.

Siepmann, D. (2016). Lexicologie et phraséologie du roman contemporain : quelques pistes pour le français et l'anglais. *Cahiers de Lexicologie*, 108, 21-41. <10.15122/isbn.978-2-406-06281-3.p.0021>

Sorba, J. (2022). *Phraséologie et genres textuels. Perspectives synchroniques et diachroniques*. Mémoire de synthèse présenté pour l'habilitation à diriger des recherches. Université Grenoble Alpes. <tel-03891983>

Tutin, A. (2010). *Sens et combinatoire lexicale : de la langue au discours*. Dossier en vue de l'habilitation à diriger des recherches. Volume 1 : Synthèse. Université Stendhal – Grenoble 3.

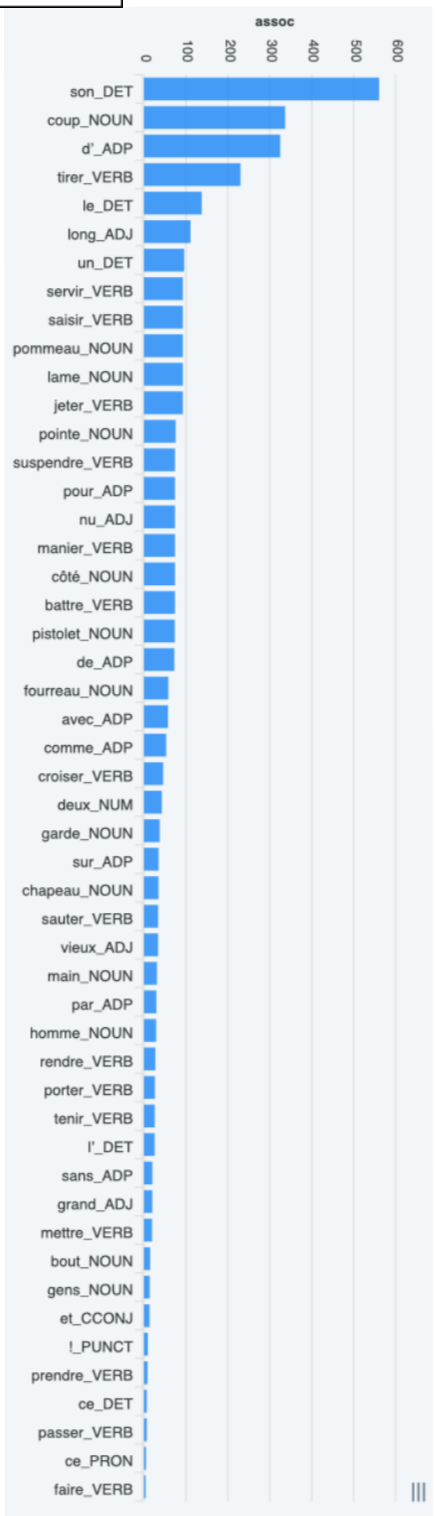
Tutin, A. (2013). Les collocations lexicales : une relation essentiellement binaire définie par la relation prédicat-argument. *Langages*, 189, 47-63. <10.3917/lang.189.0047>

Viprey, J.-M. (2006). Structure non séquentielle des textes. *Langages*, 163, 71-85. <10.3917/lang.163.0071 >

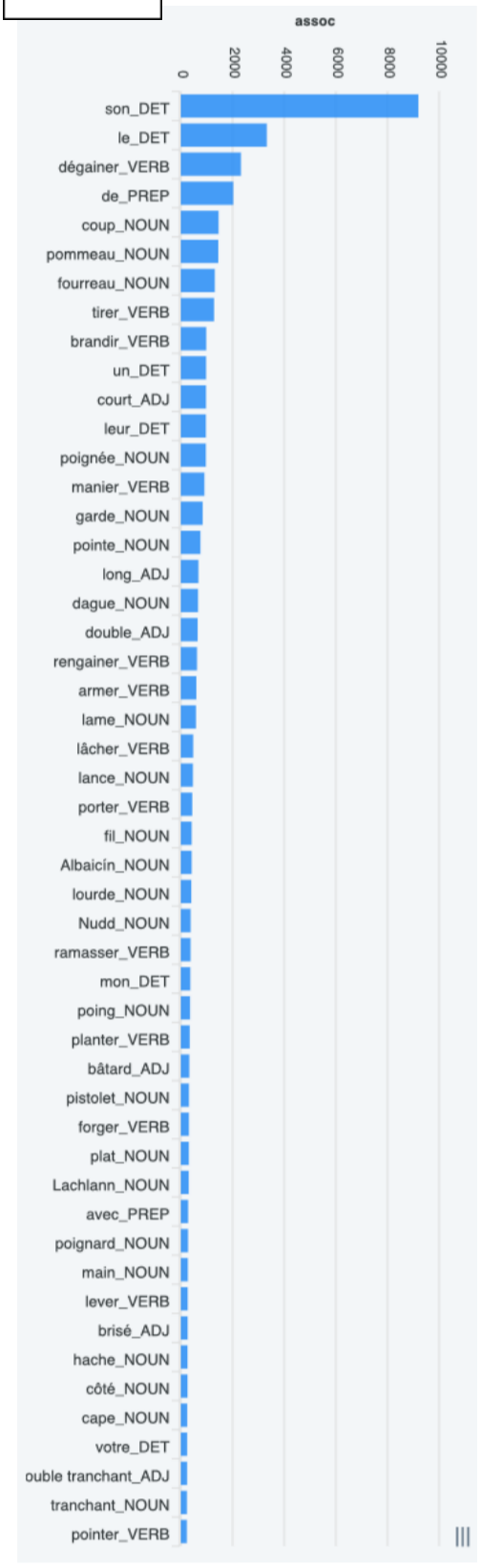
Annexes

Annexes 1 et 2. - Lexicogrammes du pivot *épée* dans le corpus phraseorom19e_fr (1) et phraseorom_fr_fr (2)

Annexe 1

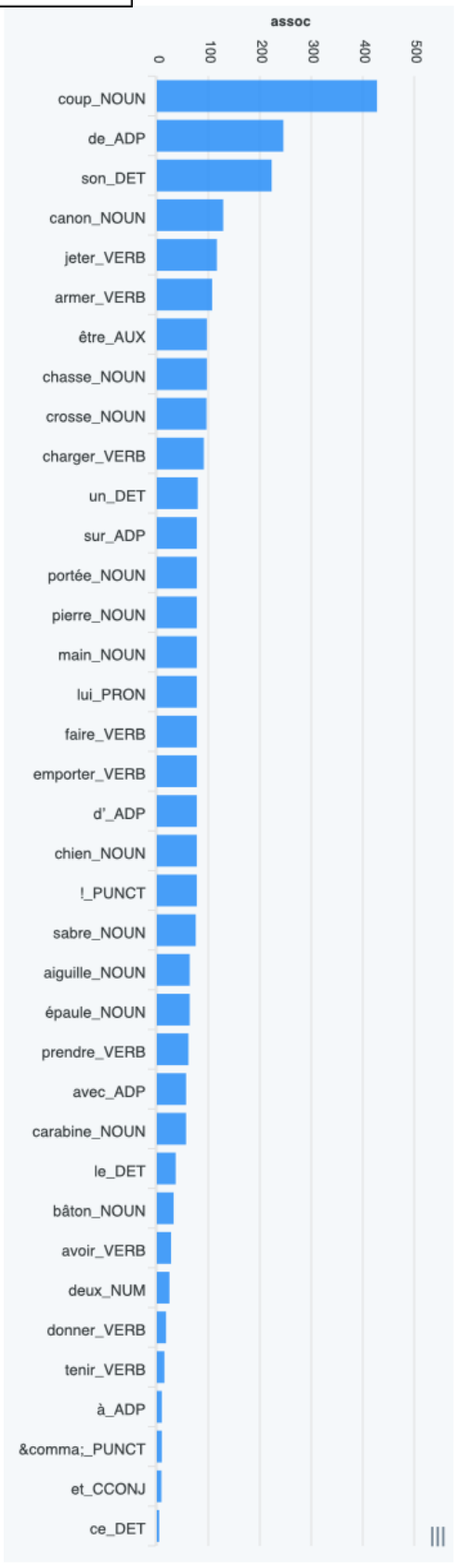


Annexe 2

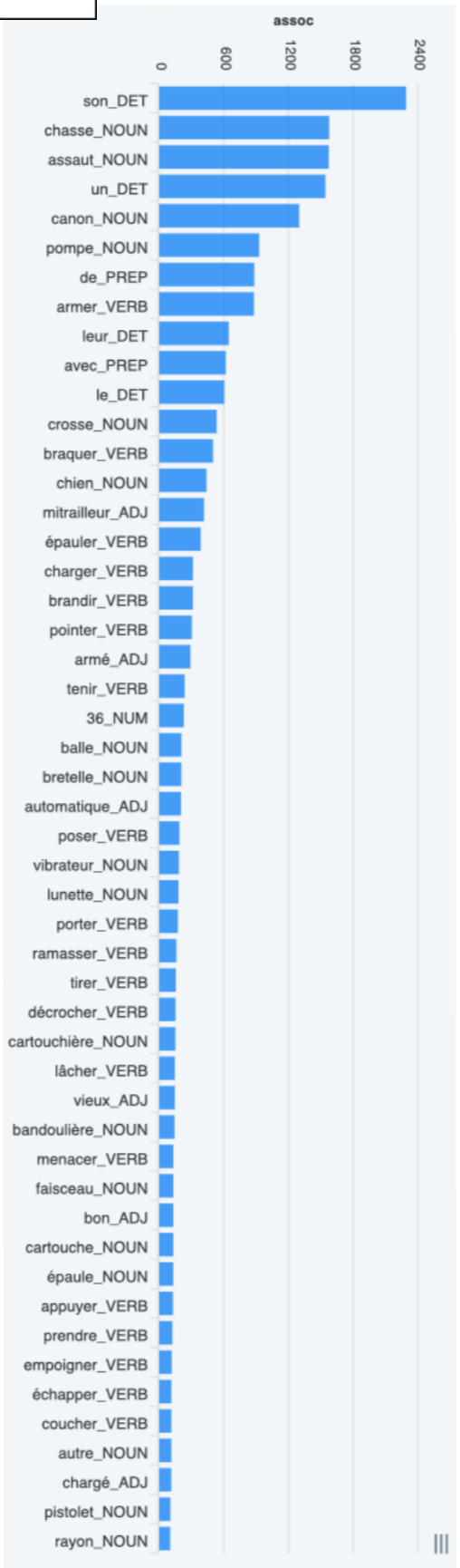


Annexes 3 et 4 : Lexicogrammes du pivot fusil dans le corpus phraseorom19e_fr (3) et phraseorom_fr_fr (4)

Annexe 3



Annexe 4



Ressources en acquisition et pathologie de l'acquisition du langage : valorisation des données sur CENHTOR

Christine da Silva-Genest [1,2], Anne-Lise Christmann [3] et Pierre Willaime [3]
[1] EA 3450 DevAH, Université de Lorraine
[2] UMR 7114 MoDyCo
[3] MSH Lorraine
christine.da-silva-genest@univ-lorraine.fr, anne-lise.christmann@univ-lorraine.fr
pierre.willaime@univ-lorraine.fr

Introduction

Le recours à l'analyse de langage spontané est de plus en plus recommandé (Bawayan & Brown, 2022; Klatte Inge et al., 2022) pour évaluer les compétences langagières des enfants. Toutefois, le manque d'outils et de données à disposition, tout particulièrement en langue française (da Silva-Genest et Masson, 2019) est un frein à la mise en place de cette pratique (Pavelko et al., 2016). Ainsi, le présent projet se donne comme objectif principal de construire une base de données d'interactions verbales adulte-enfant en situations naturelles, appelées également spontanées, et créer des outils d'évaluation fiables et efficaces. Cet objectif se concrétisera également par la mise à disposition de ces données et les ressources rattachées à celles-ci pour une exploitation diverse de la part des usagers et enrichir les bases de données existantes en acquisition et pathologie de l'acquisition du langage (Masson et al., 2021). Pour cela, nous avons souhaité déposer les données sur la plateforme numérique CENHTOR (Centre de ressourcEs Numériques des Humanités et des TerritOires) développée par la MSH-Lorraine en partenariat avec l'Inist-CNRS. Cette plateforme sert notamment à la valorisation et la diffusion des données de la recherche en Sciences Humaines et Sociales.

La présente proposition de communication exposera la construction de la page du site du projet ainsi que l'ensemble des ressources accessibles.

Corpus et méthodologie

Pour répondre à notre question de recherche générale portant sur l'évaluation du langage, nous avons constitué un corpus d'interaction adulte-enfant en situations naturelles que nous présentons ci-dessous.

Corpus

A l'heure actuelle, la population de notre étude est constituée de 83 enfants qui présentent trois profils linguistiques différents : avec un Trouble Développemental du Langage (TDL, par la suite), avec un trouble de la fluence (bégaiement) ou à développement dit typique (Cf. table 1). Les enfants sont répartis selon leur âge (de 4;6 ans à 8;7 ans) ou groupe d'âge (5, 6, 7 ou 7+ ans), leur profil linguistique, leur sexe ainsi que leur niveau scolaire (allant de la moyenne section d'école maternelle au CE1).

Population	Enfant à développement typique	Enfant présentant un TDL	Enfant présentant un bégaiement
N=	60	13	10

table 4. : table 1 :Population de l'étude

Notre corpus est donc composé de 186 interactions adulte-enfant soit environ 35 heures d'enregistrement. Nous présentons dans la section suivante la méthodologie du projet.

Méthodologie

Chaque enfant a été observé dans deux situations naturelles : une situation de jeu symbolique autour d'une maison playmobil® en interaction avec l'un des parents (le plus souvent la mère) de l'enfant et une situation de récit d'expériences personnelles avec un adulte non connu de l'enfant et membre du projet.
La situation de jeu symbolique (ou jeu libre) dure au minimum 20 minutes et celle de récit d'expériences personnelles a une durée variable dépendante de la longueur des récits des enfants. Les enfants étaient amenés à produire deux récits en racontant la journée précédant l'enregistrement ou une journée type et son meilleur ou pire souvenir.

Ces situations ont été enregistrées en audio et vidéo. Ces enregistrements ont fait l'objet d'une transcription à l'aide du logiciel CLAN (Mac Whinney, 2000). Les données des sujets sont recueillies dans le respect du Règlement Général sur la Protection des Données (RGPD).

Ces données (enregistrements et transcriptions) constituent ce que nous avons appelé les données primaires qui sont le support de l'exploitation réalisée en fonction de nos objectifs de recherche (analyses linguistiques diverses aux niveaux lexical, morphosyntaxique, pragmatique et discursif). Ce deuxième niveau de traitement permet d'obtenir un nouveau jeu de données correspondant en partie à nos résultats de recherche et que nous appelons données secondaires (cf. section suivante). L'ensemble de ces données seront accessibles via CENHTOR de manière ouverte. Cette accessibilité est réalisée de manière progressive puisque la page du projet EVALANG se construit en fonction de l'avancée du projet. CENHTOR est une ressource transdisciplinaire en sciences humaines et sociales qui permet une éditorialisation des données et qui est ainsi complémentaire à l'utilisation d'un entrepôt de données.

Résultats : Valorisation des données sur CENHTOR

Les données du projet seront mises à disposition sous différentes formes :

figure . 1 les **données primaires** : fichiers audio des interactions verbales et les transcriptions associées (au format .cha) avec métadonnées accessibles pour les utilisateurs notamment en fonction de l'âge, du niveau scolaire, des activités (jeu vs. récit d'expériences personnelles) et du profil linguistique des enfants (avec ou sans trouble développemental du langage, avec ou sans bégaiement) ;

figure . 2 les **données secondaires** constituées:

figure . 1 d'un accès aux résultats de la recherche (référentiel lexical[38], mesures d'évaluation, articles du projet, etc.);

figure . 2 d'exemples illustrant des phénomènes linguistiques fins pertinents (modalités d'étayage, production d'un énoncé complexe, énoncé disfluent vs. fluent, etc) correspondant à des extraits vidéo transcrits accompagnée d'une fiche descriptive de l'extrait.

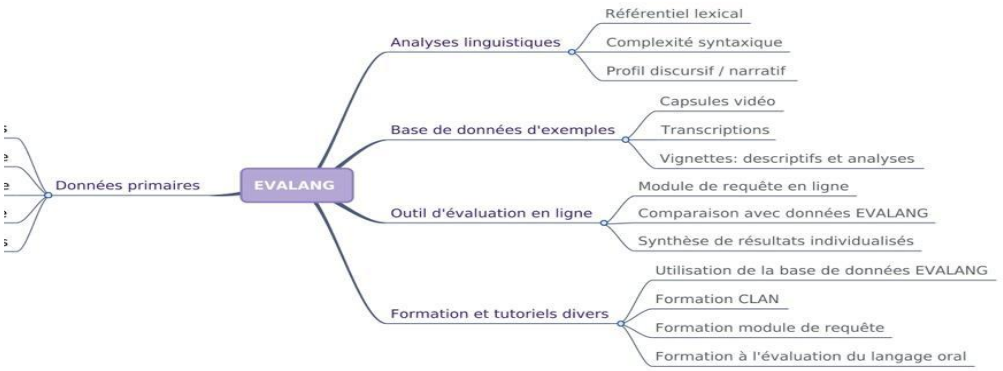
CENHTOR est la plateforme numérique de valorisation et d'exploitation des données de la recherche de la MSH Lorraine. Basée sur l'outil Omeka S, elle permet de travailler sur les données et de les visualiser. Son intérêt pour le projet réside dans la fairisation des données primaires et secondaires, leur mise à

[38] Le référentiel lexical sera présenté dans le cadre d'une communication du groupe REFLEX-EVALANG (« Création d'un référentiel lexical à partir des productions verbales d'enfants à développement typique et atypique en situation de jeu »).

disposition et leur mise en relation. Les données primaires sont tout d'abord ingérées dans Omeka S avec au passage un alignement sur des référentiels d'autorité lorsque cela est possible et, plus largement, des actions visant à renforcer la fairisation des données (Aucagne et al., 2022 "Vademecum pour la réutilisabilité des données"). Les données sont ensuite présentées de manière à permettre d'effectuer des recherches dans le corpus. La plateforme intègre des filtres permettant aux chercheurs de pouvoir cibler précisément un ou plusieurs cas d'étude. Les différents types de données correspondant aux différentes étapes du projet sont également mises en relation de manière à pouvoir passer d'une donnée secondaire aux données primaires liées.

La communication présentera ces données et la manière d'y accéder sur la plateforme. Parmi les données secondaires, nous reviendrons sur la mise à disposition d'un référentiel lexical (constitué des données portant sur la diversité lexicale des enfants ainsi que les formes produites en situation de jeu en fonction de l'âge, du niveau scolaire, du sexe) et sur la base de données d'exemples.

A terme, nous souhaitons également mettre à disposition des outils d'évaluation via un module de requête en ligne ainsi que des formations et divers tutoriels portant sur les thématiques du projet (*e.g.* évaluation du langage, langage oral, transcrire des productions verbales, analyse du langage spontané, etc.). Ces capsules sous forme de vidéo ont pour objectif de servir la formation initiale d'étudiant·es issu·es de diverses disciplines (sciences du langage, orthophonie, psychologie développementale) ainsi que la formation continue de chercheur·es et de professionnel·les de santé. La figure ci-dessous présente la diversité des données que l'on souhaite mettre à disposition à diverses échéances (courte, moyenne et longue). La plateforme CENHTOR présente l'intérêt de pouvoir suivre la dynamique du projet et actualiser le site en fonction.



Pour conclure, le projet souhaite ainsi contribuer au développement des savoirs et des outils des humanités numériques, mettre à disposition et valoriser les données issues de la recherche sous différentes formes servant tant l'enseignement et la recherche que la clinique.

Références bibliographiques

Aucagne, J., Bordry, M., Desiles, C., Filoche, F., Garcia-Fernandez, A., Elisabeth, G., Koskas, C., Patat, G., Walter, R., & Willaime, P. (2022). Vademecum pour la réutilisabilité des données [Research Report]. Consortium CAHIER - Huma-Num. https://hal.univ-grenoble-alpes.fr/hal-03630095

Bawayan, R. & Brown, J.A. (2022). Language sample analysis consideration and use: a survey of school-based speech language pathologists. *Clinical Archives of Communication Disorders* 7(1):15-28.

da Silva-Genest, C. & Masson, C. (2019). Corpus et pathologies du langage : du recueil à l'analyse de données pour une linguistique clinique et appliquée, *Corpus* [En ligne], 19 | 2019, mis en ligne le 01

janvier 2019, consulté le 02 juin 2023. URL : http://journals.openedition.org/corpus/4374 ; DOI : https://doi.org/10.4000/corpus.4374

Klatte Inge S., van Heugten Vera, Zwitserlood Rob, & Gerrits Ellen. (2022). Language Sample Analysis in Clinical Practice : Speech-Language Pathologists' Barriers, Facilitators, and Needs. Language, Speech, and Hearing Services in Schools, 53(1), 1-16. https://doi.org/10.1044/2021_LSHSS-21-00026

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ (Lawrence Erlbaum Associates).

Masson, C., da Silva-Genest, C., Canut, E. & Caët, S. (2021). Usages des corpus oraux pour l'étude de l'acquisition du français langue maternelle. In C. Benzitoun & M. Rebuschi (Eds), *Les corpus en sciences humaines et sociales*. Nancy : Collection MSHL, 15-48.

Pavelko, S. L., Owens, R. E., Jr., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school- based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools*, 47(3), 246–258. DOI: https://doi.org/10.1044/2016_LSHSS-15-0044

Création d'un référentiel lexical à partir des productions verbales d'enfants à développement typique et atypique en situation de jeu

Christine Da Silva-Genest [1], Loïc Liégeois[2], Caroline Masson[3], Christophe Benzitoun[4], Marine Le Mené Guigourès[5]
[1] Laboratoire DevAH, Université de Lorraine
[2] Laboratoire LLF, Laboratoire CLILLAC-ARP, Université de Paris Cité
[3] Laboratoire CLESTHIA, Université Sorbonne Nouvelle
[4] Laboratoire ATILF, Université de Lorraine
[5] Laboratoire CRBML, Université du Québec à Montréal
christine.da-silva-genest@univ-lorraine.fr, loic.liegeois@u-paris.fr, caroline.masson@sorbonne-nouvelle.fr,
christophe.benzitoun@univ-lorraine.fr, le_mene_guigoures.marine@uqam.ca

Introduction

Quand il s'agit d'évaluer le langage d'enfants en contexte clinique, les compétences lexicales sont le plus souvent évaluées par des tâches de dénomination et de désignation d'images issues de batterie de tests de langage (Thibaut et al., 2001). Si la plupart des orthophonistes considèrent ces tests utiles dans le cadre de la prise en charge de jeunes patients, ce type d'épreuves est aussi critiqué (Rondal, 2003; Thibaut et al., 2001). Les raisons sont diverses et portent le plus souvent sur les caractéristiques psychométriques des tests (Bignon et al., 2021), leur capacité à discriminer de manière précise un trouble développemental (Spaulding et al., 2006), ou encore la pertinence des items testés au sein des épreuves (da Silva-Genest et al., 2020 ; Rondal, 2003). En outre, ces tâches formelles ne tendent pas à rendre compte de l'ensemble des compétences des enfants. En effet, elles évaluent les déficits linguistiques en faisant référence à une seule et unique courbe développementale et, au niveau lexical, à la connaissance ou la non connaissance des items testés.

Face à ces diverses problématiques, l'analyse du langage spontané est une méthode d'évaluation de plus en plus recommandée (Klatte Inge et al., 2022) pour compléter les données issues des tests ou lorsqu'une évaluation formelle est impossible. Cette méthode permet d'apprécier davantage les compétences réelles des enfants et la façon dont ils les mobilisent. De récentes études ont déjà pu montrer la pertinence et l'efficacité de cette méthode (Imgrund et al., 2019) ainsi que ses nombreux avantages (*e.g.* situation naturelle, pas d'effet test-retest). Cependant, cette méthode est encore peu pratiquée et n'est pas réalisée de façon systématique (Pavelko et al., 2016). De nombreux freins sont relevés tels que la méthodologie utilisée pour le recueil, la difficulté de traitement des données (transcription, annotation, analyses linguistiques, etc.) et l'interprétation des résultats. Pour pallier ces difficultés, les chercheurs et cliniciens tentent de proposer des alternatives et des solutions adaptées comme limiter le temps d'enregistrement et d'analyse ou le nombre d'énoncés (Guo & Eisenberg, 2015) et avoir recours à l'utilisation d'outils d'analyse automatique efficaces et pertinents (da Silva-Genest & Masson, 2017 ; Guo et al., 2015). Toutefois, la littérature dans ce domaine reste encore limitée, et ce tout particulièrement en langue française. Ainsi, nous souhaitons pallier ce manque en apportant de nouvelles connaissances sur le développement linguistique d'enfants âgés de 4;6 à 8 ans en situations naturelles et en créant des outils d'évaluation fiables et efficaces. Dans le cadre de la présente proposition de communication, nous nous focaliserons sur les compétences lexicales des enfants.

Corpus et méthodologie

Corpus

Notre corpus est constitué de 59 interactions parent-enfant en situation de jeu autour d'une maison contenant des personnages et des objets que les participants peuvent manipuler et mettre en scène comme ils le souhaitent en imaginant une histoire.

La population de l'étude est donc constituée de 59 enfants : 38 enfants à développement typique âgés de 4;7 ans à 7;5 ans, 11 enfants présentant un Trouble Développemental du Langage oral âgés de 5;2 ans à 8;7 ans et 10 enfants présentant un bégaiement âgés de 4;3 ans à 6;10 ans. De plus, différents paramètres ont été considérés à savoir : le profil linguistique des enfants (avec ou sans Trouble Développemental du Langage ; avec ou sans bégaiement), leur niveau scolaire (maternelle ou élémentaire) et le sexe (fille / garçon).

Méthodologie

Toutes les interactions verbales ont été enregistrées en audio et vidéo et durent au minimum 15 minutes. Ces interactions ont fait l'objet d'une transcription à l'aide du logiciel CLAN (MacWhinney, 2000).

L'analyse des données permettra de constituer un **référentiel lexical** composé d'une part d'une mesure de diversité lexicale des productions des sujets et d'autre part, d'une liste des lexèmes produits par les enfants accompagnés des informations concernant leur catégorie morphosyntaxique et leur fréquence d'usage en situation de jeu. La mesure de la diversité lexicale des productions des sujets (Voc-D) sera obtenue au moyen de la commande VOCD (Bernstein Ratner et MacWhinney, 2016; Malvern & Richards, 1997) disponible dans CLAN.

La commande VOCD a fait l'objet de nombreuses études qui relèvent son intérêt et la présentent comme une mesure efficiente et robuste (voir par exemple McCarthy et Jarvis, 2010, pour une revue des méthodes de calcul de la diversité lexicale). Par opposition au Type-Token Ratio (TTR) qui n'est pas recommandé pour un usage clinique, VocD est considéré comme plus fiable et efficace (Yang et al., 2022).

Résultats

Le référentiel lexical sera présenté en tenant compte des deux aspects considérés à savoir l'indice de diversité lexicale et les formes produites par les enfants.

Au niveau développemental, les résultats provenant de l'analyse de la diversité lexicale via la commande VocD rendent compte d'une plus grande richesse pour les enfants plus âgés (6 et 7 ans *vs*. 5 ans) et scolarisés en école élémentaire (*vs*. maternelle). En revanche, il ne semble pas y avoir de différence au niveau lexical entre la population féminine et masculine. En outre, lorsqu'on compare les différents profils linguistiques, les enfants ayant un TDL présentent une diversité lexicale moins importante que les enfants tout-venant et bègues de même âge chronologique, et ce tout particulièrement pour la catégorie des verbes. Le score de VocD obtenu par les enfants bègues est plus élevé que celui de la population TDL et tend à s'approcher de celui des enfants tout-venant, ce qui confirme certaines données de la littérature (Luckman et al., 2020).

La constitution du référentiel lexical propre à la situation de jeu libre met également en évidence un vocabulaire commun aux enfants qui est fortement influencé par l'activité (le jeu) et le support (une maison). En effet, on relève par exemple des champs sémantiques liés au repas (*table*, *tasse*, *manger*) ainsi qu'aux objets et/ou personnages présents (*poussette*, *jumeaux*, *toit*). Certains lexèmes sont relativement fréquents dans le discours des enfants (*e.g. aller, tomber, enfant, école, chaise, chat*) alors que d'autres sont plus rares (*e.g. bouquet, brûler, coq, fabriquer, faufiler, fauteuil*). Parmi les mots fréquents, toutes les catégories morphosyntaxiques sont représentées : les noms (*voiture, truc, train, lit,*

sac, soir, école, fille, parent, enfant, …), verbes (*aller, voir, venir, se coucher, tomber, tenir, rentrer, prendre, vouloir,* etc.), adjectifs (*petit*), pronoms (*qui, que, elle,* etc.), adverbes (*encore, après, trop, très, tout)*, prépositions (*dans, sur, pour, avec*), conjonctions (*que, pour que, parce que*) et déterminants.

Le référentiel lexical créé dans le cadre du présent projet est intéressant à plus d'un titre. En effet, il permet à la fois de relever les lexèmes produits ainsi que leur fréquence par les enfants en situation de jeu et en interaction avec un adulte et de dégager de manière complémentaire une mesure d'évaluation lexicale qui différencie les profils linguistiques et rend compte des courbes développementales.

Par ailleurs, ce référentiel permet de mettre en évidence l'apport du jeu libre ou symbolique. En effet, la caractérisation du lexique produit par les enfants dans cette situation montre que cette dernière peut être propice pour travailler certaines formes fréquentes ou rares liées à la temporalité (*soir*, *après-midi*), à la personne (*moi*, *toi*, *je*, *te*, *lui*), certains champs lexicaux ou catégories morphosyntaxiques telles que les pronoms relatifs, les déterminants ou les prépositions. Le jeu libre est donc une activité riche qui peut être explorée et exploitée de diverses façons que ce soit en termes d'évaluation ou d'intervention.

Ce référentiel lexical sera mis à disposition et diffusé sur la plateforme CENHTOR (Centre de ressourcEs Numériques des Humanités et des TerritOires) développée par la MSH-Lorraine en partenariat avec l'Inist-CNRS. En outre, au-delà de cette mise à disposition, nous souhaitons également accompagner les utilisateurs de ce référentiel par la mise en ligne de modules de formation destinés à un public divers (étudiant·es de sciences du langage, orthophonie, psychologie, clinicien·nes, chercheur·es) dans le but de les former à l'évaluation de l'analyse outillée du langage spontané. En effet, comme nous l'avons noté en introduction, l'analyse du langage spontané est fortement recommandée mais cette méthode est particulièrement complexe à prendre en main par les professionnel·les de santé notamment à cause du manque de formation à ce type d'analyse (Klatte Inge S. et al., 2022) qui requiert une expertise linguistique (Rondal, 2003).

Références bibliographiques

Bernstein Ratner N. & MacWhinney B. (2016). Your Laptop to the Rescue: Using the Child Language Data Exchange System Archive and CLAN Utilities to Improve Child Language Sample Analysis. *Semin Speech Lang*. 37(2), 74-84.

Bignon, M., Gamot, L., Lemaitre, M.-P., & Macchi, L. (2021). Cotation des tests francophones de langage oral et écrit chez l'enfant : Quelques recommandations à l'usage des orthophonistes. *Glossa*, 131, 1-32.

Da Silva-Genest, C. & Masson, C. (2017). Apport de la linguistique de corpus à l'étude des situations cliniques : utilisation de ressources écologiques pour évaluer les pratiques professionnelles. *Studii de Linguistica,* 7, 89-112.

Da Silva-Genest, C., Liégeois, L., Masson, C., Benzitoun, C. et Le Mené, M. (2020). *Rabot*, *sécateur*, *téléphérique* : interroger la pertinence des choix lexicaux dans les outils d'évaluation du langage en orthophonie. *Lidil*, 62. DOI : https://doi.org/10.4000/lidil.8422

Guo, L.Y. & Eisenberg, S. (2015). Sample Length Affects the Reliability of Language Sample Measures in 3-Year-Olds: Evidence From Parent-Elicited Conversational Samples. *Language, Speech, and Hearing Services in Schools* 46: 141-153.

Imgrund, C. M., Loeb, D. F., & Barlow, S. M. (2019). Expressive Language in Preschoolers Born Preterm: Results of Language Sample Analysis and Standardized Assessment. *Journal of speech, language, and hearing research*, 62(4), 884–895.

Klatte Inge S., van Heugten Vera, Zwitserlood Rob, & Gerrits Ellen. (2022). Language Sample Analysis in Clinical Practice : Speech-Language Pathologists' Barriers, Facilitators, and Needs. *Language, Speech, and Hearing Services in Schools*, 53(1), 1‑16.

Luckman C., Wagovich S. A., Weber C., Brown B., Chang S. E., Hall N. E., Bernstein Ratner N. Lexical diversity and lexical skills in children who stutter. *Fluency Disord*. 2020, 63. DOI : 10.1016/j.jfludis.2020.105747.

Malvern, D. D. & Richards, B. J. (1997). A new measure of lexical diversity. In Ryan, A. & Wray, A. (eds), *Evolving Models of Language*. Clevedon : Multilingual Matters. 58-71.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. Language Testing, 24(4), 459-488.

McCarthy, P.M., Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods,* 42, 381-392.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Lawrence Erlbaum Associates.

Pavelko, S. L., Owens, R. E., Jr., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school- based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools*, 47(3), 246-258.

Rondal, J.A. (2003). *L'évaluation du langage*. Mardaga.

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impariment : Is the low end of normal always appropriate. *Language, Speech and Hearing Services in the Schools*, 37, 61-72.

Thibaut, J.P., Gregoire, J. & Lion, P. (2001). Sur les difficultés de l'évaluation du lexique : une perspective développementale. *Glossa* 76, 52-61.

Yang J. S., Rosvold C., Bernstein Ratner N. (2022). Measurement of Lexical Diversity in Children's Spoken Language: Computational and Conceptual Considerations. *Front Psychol*. 13 DOI: 10.3389/fpsyg.2022.905789.

Des « petites phrases » à la phrase : constitution et exploitation d'un corpus de discours politico-médiatiques

Damien Deias
Centre de recherche sur les médiations (CREM), Université de Lorraine
damien.deias@univ-lorraine.fr

Introduction

« *Mon véritable adversaire, c'est le monde de la finance* » (François Hollande, 22/01/2012), « *La République, c'est moi !* » (Jean-Luc Mélenchon, 16/10/2018) « *Quand j'entends le mot "violence policière", je m'étouffe* » (Gérald Darmanin, 28/07/2020) « *Les non-vaccinés, j'ai très envie de les emmerder* » (Emmanuel Macron, 04/01/2022), « *Les gens ne connaissent pas bien l'économie* » (Bernard Arnault, 26/01/2023)… Les productions discursives détachées, reproduites et nommées « petites phrases » par les journalistes semblent ponctuer la vie politique et médiatique. Elles sont parfois considérées comme des événements majeurs, saturant provisoirement l'espace médiatique, traduites dans la presse étrangère, phénomène que Maingueneau nomme la « panaphorisation » (Maingueneau, 2012 : 89).

Elles constituent donc, pour plusieurs raisons, un objet de choix pour le linguiste, interrogeant différents champs des sciences du langage. Le rapport de l'objet à sa dénomination, la co-construction de l'objet entre les acteurs politiques et médiatiques, sa forme syntaxique, sa capacité à circuler et les modifications de l'énoncé qui l'accompagnent, sa capacité à générer quantité de discours en rapport à sa brièveté ou bien encore sa force illocutoire et perlocutoire sont autant de questions et de problèmes qui en font en objet particulièrement riche mais également délicat à appréhender. L'objectif de cette communication est de présenter l'analyse de cet objet discursif complexe par l'élaboration d'un corpus ouvert, hétérogène et plurisémiotique.

Bien que des fragments discursifs qualifiés de « petites phrases » soient visibles dans la presse écrite quotidienne dès les années 1970, les linguistes et chercheurs en Sciences de l'information et de la communication ne s'en sont emparés que récemment. Outre l'article de Brasart (1994), le dossier coordonné par Krieg-Planque et Ollivier-Yaniv (2011) a permis d'envisager l'étude systématique de l'objet, et le dossier coordonné par Boyer et Gaboriaux (2018) de mettre en lumière les stratégies discursives qui l'animent. Maingueneau a fourni des outils indispensables pour comprendre l'énonciation des « phrases sans texte secondaires », c'est-à-dire les phrases « *qui ont été prélevées sur un texte source* » (Maingueneau, 2012 : 8). Nous avons, pour notre part, réalisé une étude exhaustive et à visée systématique de l'objet « petite phrase », à partir d'un large corpus de données collectées. Notre poster abordera la problématique de l'élaboration d'un corpus qui prend en compte la particularité de la co-construction d'une petite phrase, produite à l'oral par un acteur politique, détachée par un acteur médiatique. Nous aborderons également la problématique de sa circulation à travers différents espaces textuels et discursifs qui provoque des modifications et l'apparition de différentes formulations. Le traitement des données issues de ce corpus a permis d'obtenir des résultats relatifs aux différentes dimensions de l'objet. Nous présenterons sur ce poster les résultats qui concernent la dimension syntaxique de l'objet, permettant une réflexion originale sur la notion de « phrase ».

Problématique et problèmes d'un corpus de « petites phrases » politiques

L'objet discursif « petite phrase »

Détacher une phrase d'un texte ou d'un discours n'est pas une pratique récente. Il en va de même pour l'action consistant à anticiper le détachement d'une phrase d'un discours que Maingueneau nomme la « surassertion » (Maingueneau, 2004). Le théâtre classique français par exemple connait cette pratique et des énoncés détachés du *Cid* de Corneille circulent aujourd'hui en qualité de citation ou de « particitation », sans que ne soient nommés ni l'auteur ni l'œuvre. La dénomination « petite phrase », quant à elle, a été forgée par les journalistes et spécialistes des médias.

Les petites phrases se distinguent par plusieurs critères. Il s'agit d'énoncés appartenant au domaine politique, détachés par des journalistes pour entrer dans un processus de circulation médiatique. Il y a donc co-construction, et ce même si le rôle des acteurs politiques et des acteurs médiatiques n'est pas symétrique. Ces énoncés sont très majoritairement oraux. L'étape du détachement correspond donc également à une étape de transcription et d'inscription d'un énoncé oral dans l'ordre de l'écrit. Cette étape d'intégration de la petite phrase dans le discours médiatique demande donc une opération de sélection d'énoncés dans le discours politique, et peut impliquer des modifications et adaptations de ces énoncés par les acteurs médiatiques.

Prenons pour exemple l'énoncé suivant d'Emmanuel Macron, devenu une petite phrase retentissante, reproduite de la sorte dans *Ouest France* : « *Les salariées de Gad sont pour beaucoup illettrées.* » (17/09/2014). La fragment correspondant, extrait d'un entretien accordé à la radio Europe 1, n'est pas identique : « *Il y a dans cette société une majorité de femmes, il y a qui sont pour beaucoup illettrées.* ». Il n'est pas question, ici de parler d'intention de falsification. Cependant, force est de constater que l'énoncé a été modifié pour devenir une petite phrase, et que notre corpus montre des récurrences dans ces modifications, assimilables à une routine journalistique. La question de la forme syntaxique de ces énoncés et de la progression thématique est alors l'une des problématiques centrales dans la description linguistique de cet objet.

Un corpus nécessairement ouvert et hétérogène

Un corpus ayant pour objectif l'étude des petites phrases politiques ne peut être qu'un corpus ayant pour objectif l'étude du phénomène des petites phrases. Il ne peut, à la manière de certains corpus rassemblant d'autres productions discursives brèves comme les proverbes (Sevilla Munoz, 2000), se contenter de lister les énoncés circulants que constituent les petites phrases. Il se doit, ainsi que dans la démarche envisagée par Cislaru et Sitri, d'intégrer les « conditions de production » et les « extérieurs » des petites phrases :

> En AD, le corpus n'est pas seulement construit, comme dans la plupart des domaines de la linguistique, en fonction d'un objectif de recherche ; il est, par ailleurs, contextualisé et mis en relation avec des « conditions de production », avec des pratiques sociales, plus largement avec des extérieurs qui le déterminent. (Cislaru & Sitri, 2012 : 61)

Ce faisant, nous avons collecté des données permettant de rendre compte en amont de l'acte de détachement des petites phrases, par la collecte de discours sources, et en aval de leur intégration à des discours médiatiques et de leur circulation, par la collecte d'articles de presse, de publications sur les réseaux sociaux numériques, de mèmes etc. Nous avons ainsi collecté 181 petites phrases ainsi que leur discours source. Nous avons ensuite mené un travail de collecte de discours dans lesquels ces petites phrases ont circulé. Pour cela, nous avons eu recours à la base de donnée Europresse, et nous sommes également rendu directement sur les principaux réseaux sociaux numériques, et en particulier Facebook, Twitter et TikTok. Cette démarche nous a permis de caractériser et définir précisément l'objet, par l'utilisation de la notion de « prototype ». Il nous a également permis d'identifier les différentes modifications syntaxiques des énoncés, en fonction notamment du genre de discours où ils sont reproduits. Cette démarche s'inscrit dans les tendances françaises de l'analyse du discours

(Maingueneau & Charaudeau, 2002 : 202), accordant une place prépondérante à l'énonciation, et visant à articuler l'étude de phénomènes linguistiques sur des dispositifs de communication.

Résultats

De la « petite phrase » à la phrase

Nous choisissons de présenter dans cette communication les résultats portant sur la forme syntaxique des petites phrases. La dénomination « petite phrase » est le résultat d'un figement à l'opacité relative. Elle désigne des objets discursifs qui se rapportent au domaine médiatique. Ces objets discursifs correspondent le plus souvent à la notion grammaticale de « phrase ». Une fois ce constat opéré, l'exploitation de notre corpus nous conduit à mettre en lumière ce que nous avons nommé des « calibres syntaxiques » auxquels correspondent la plupart des petites phrases. Nous remarquons ainsi la prévalence de petites phrases à structure bipartite, notamment issues de réarrangements thématiques (« *Mon adversaire, c'est le monde de la finance* », Hollande, *Le Monde*, 13/05/2017), des constructions à constituant détaché (« *La première usine qu'il faut faire en France, c'est une usine à couilles !* » J-M Le Pen, 2012) ou bien encore les constructions clivées (« *il y a des fessées qui se perdent parfois* » (M Alliot-Marie, 26/07/2014). La prévalence de ces constructions et des réarrangements thématiques éventuels opérés par les journalistes peut être en partie expliquée par la grammaire de la période de Berrendonner et du groupe de Fribourg (2009).

De la phrase à la période

La notion de « phrase » est très discutée en sciences du langage, et parfois remise en cause pour d'autres modèles (Benzitoun, 2011), (Debaisieux, 2013). A l'aune de notre corpus de petites phrases, nous retenons en particulier le modèle de Berrendonner en ce qu'il propose, par les notions de « clause » et de « période », une intégration des structures syntaxiques, prosodiques et pragmatiques. Il est alors remarquable de constater le rendement significatif de deux structures macro-syntaxiques, que nous illustrons ici avec deux énoncés de Nicolas Sarkozy :

> Préparation > Action : « (*Le drame de l'Afrique,*) > (*c'est que l'homme africain n'est pas assez entré dans l'histoire.*) »

6 Action + confirmation : « (*Je veux que partout dans le monde, les opprimés, les femmes martyrisées, les enfants emprisonnés ou condamnés au travail, sachent qu'il y a un pays dans le monde qui sera généreux pour tous les persécutés*,) (*c'est la France*) »

L'intérêt de tels résultats est double à nos yeux. Ils nous renseignent sur le fonctionnement discursif et syntaxique des petites phrases, et dans un même temps permettent d'appréhender la représentation spontanée de la phrase.

Références bibliographiques

Benzitoun, C. *et al.* (2011), *tu veux couper là faut dire pourquoi.* Propositions pour une segmentation syntaxique du français parlé. *2ème Congrès Mondial de Linguistique Française*, 139.

Berrendonner, A. (2009). *Grammaire de la période*. Peter Lang.

Brasart, P. (1994), Petites phrases et grands discours (Sur quelques problèmes de l'écoute du genre délibératif sous la Révolution française). *Mots. Les langages du politique, 40*, 106-112.

Boyer, H. & Gaboriaux, C. (2018), Splendeurs et misères des *petites phrases*. *Mots. Les langages du politique, 117*, 9-17.

Cislaru, G. & Sitri, F. (2012), De l'émergence à l'impact social des discours : hétérogénéités d'un corpus. *Langages, 187(3)*, 59-72.

Debaisieux, J.-M. (2013). *Analyses linguistiques du corpus : subordination et insubordination en français.* Hermès-Lavoisier.

Krieg-Planque, A. & Ollivier-Yaniv, C. (2011), Poser les « petites phrases » comme un objet d'étude. *Communication & Langages, 168*, 17-22.

Maingueneau, D. & Charaudeau, P. (dir.) (2002), *Dictionnaire d'analyse du discours*, Seuil.

Maingueneau, D. (2004). Citation et surassertion. *Polifonia, 8* ; 1-22.

Maingueneau, D. (2012). *Les phrases sans texte.* Armand Colin.

Sevilla Munoz, J. (2000). Les proverbes et phrases proverbiales français, et leur équivalences en espagnol. *Langages, 34(139)*, 91-106.

The Use of Corpus Consultation in Translation Revision

Tokdemir Demirel Elif [1], Öztürk Muzaffer [2] , Toprakçı Yağmur Sude [3], et Çiçek Zeynep Betül [4]
[1, 2, 3, 4],Department of Translation and Interpretation, Kırıkkale University, Kırıkkale/Turkey
demirel@kku.edu.tr, muzafferozturk_@outlook.com, yagmursudetoprakci@gmail.com, zynp.btlcck@gmail.com

Introduction

This study will report the findings of an ongoing project supported TUBITAK (The Scientific and Technological Research Council of Turkey) 2209-A - Research Project Support Programme for Undergraduate Students. Today, translation technologies have come to a very advanced point and machine translation methods are also very advanced. Despite this, the role of well-trained translators in the field of translation is undeniable. Translation is a job that needs to be done in detail and there is a possibility of errors in translation works. Written sources and dictionaries on grammar are generally used to correct these errors, but a resource that is thought to be productive is the consultation of a corpus as a reference tool. In linguistic terms, a corpus can be defined as "a collection of spoken and written texts organized by genre and coded according to various discourse elements". (Biber et al., 1999, p4). Today, the fact that corpora is online and free resources has made them easily accessible and many researchers refer to corpora for language research. The use of corpus has also become widespread in research and practice in the field of translation. For this reason, in this study, it will be examined how effective corpus consultation can be in the editing of translation

The aim of this research is to investigate the extent to which errors in different categories that occur in student translations can be revised by students by using corpus consultation and what kind of differences it will show in translation quality. Both quantitative and qualitative methods will be used in the research. Two volunteer student groups will be formed as control and experimental groups. Both groups will be asked to translate a section from the children's book Little Prince (de Saint-Exupéry, A., 1943) and the errors of these translations will be categorized and determined by the project team. After the errors are identified, the experimental group will be given training on the use of corpus, and they will be asked to correct the translation errors, and these corrections will be compared with the error corrections made by the control group with traditional methods. As a result of these comparisons, answers to the research questions of the study will be sought. In addition, a survey will be conducted with participant student groups to measure attitudes towards corpus consultation. Based on the results of the study, a study model will be designed that the students of the Department of Translation and Interpretation can use in translation revision applications.

In order to understand the importance of the subject, we must first understand what a corpus is and the importance of post-translational editing. The concept of corpus is defined by Biber (1998) as regular combined natural texts, and Kennedy (1998, p.3) as a collection of texts collected in an electronic database. Thanks to the fact that the corpus is available in the electronic environment, we can easily learn the information in corpora that may escape attention when the language is studied only by observation. There are many corpus studies all over the world. Major projects for corpus development began in the 1980s and early 1990s. Some of the major corpus projects leading corpus linguistics projects are COBUILD Corpus, Bank of English Corpus, ICE Corpus, Wellington Corpus, and Helsinki Corpus. Among these corpus projects, some still stand out and continue to evolve, such as BNC, and are widely used by researchers and language learners.

When it comes to post-translation editing, every translation, be it machine or human translation, contains errors, especially translations by students of the translation department. There may be many things that are overlooked during the translation, but these errors can be seen more easily with a holistic view of the

text afterwards. The errors mentioned are not just about the lack of grammar. The way the local people use the language has been effective in the formation of all languages. For this reason, the person/translator who learns a second language does not have the cultural knowledge of the local speakers of that language, and accordingly, the deficiencies in the knowledge of collocations in the language may reduce the translation quality. Since it will take a long time to learn cultural knowledge, the use of corpus is considered very important in this research. Considering that incomplete or overlooked mistakes such as syntax and collocation on the texts collected in electronic environment and easily accessible will be seen more easily, this research will be carried out with the foresight that if positive data are collected as a result of this research, a practical training on the use of corpus will increase the translation quality and make the translator's job easier.

The aim of this project is to enable students who have completed their translations to autonomously detect and correct their translation errors using corpus consultation, and to minimize translation errors in the long run. Bulut (2015, p.68) states that the translation will be a comedy element when it turns to areas of wrong meaning. In this context, minimizing errors will have a significant positive effect on the perception of translation. This project aims to help students correct these translation errors faster and improve the quality of their translations with the compilation method. It is also aimed to contribute to translation education through the correction process and to increase the awareness level of the students.

Corpus and Methodology

Corpus

In the study a corpus of student translations of sections of the children's book « Little Prince » has been compiled. The translations were done from Turkish to English. The student translations were done during two in-class sessions guided by two translation task sheets prepared by the researchers. A total of 30 students participated in the translation phase of the study and consent was taken from the students participating in the study. At the translation phase, sections from the first part of the book were aligned in a table format and students were provided space to write their translations sentence by sentence. Part of the tasksheet can be seen in Figure 1 below.

The translations were done by hand by the students in order to endure that students do not use machine translation ; however, the students were allowed to use online or printed dictionaries. The translations were then transferred into the computer by the project team and compiled into a small corpus named as LPCorp (Little Prince Corpus) consisting of approximately 15.000 words. The student translation corpus was then analyzed and coded for translation errors. A translation typology was created in order to code the errors by referring to the literature (Munday, 2016).

LITTLE PRINCE TRANSLATION TASK

PART 1

SOURCE TEXT	TARGET TEXT
Altı yaşındayken bir gün, balta girmemiş ormanlar üstüne yazılmış "Yaşanmış Öyküler" adlı bir kitapta müthiş bir resim görmüştüm.	
Bir hayvanı yutmakta olan bir boa yılanını gösteriyordu. Resmin kopyası işte yukarıda.	
Kitapta şöyle deniyordu: "Boa yılanları avlarını çiğnemeden olduğu gibi yutuverirler.	
Sonra da yerlerinden kımıldayamaz, sindirimleri için gerekli altı ayı uyumakla geçirirler."	

figure . 66 Translation Task 1

Error Categories		Error Codes
Linguistic Errors	Morphological	[L-M]
	Syntactic	[L-S]
	Collocations	[L-C]
	Spelling	[L-S]
Comprehension Errors	Misunderstanding of Lexis[U]	[C-ML]
	Misunderstanding of Syntax	[C-MS]
Translation Errors	Distorted Meaning	[T-DM]
	Additions	[T-A]
	Omissions	[T-O]
	Inaccurate Renditions of Lexical Items	[T-LI]
Unidentified	Unidentified error	[U]

table 5. : The Typology of Errors

Methodology

In this project, both qualitative and quantitative methods were used. Document review, translation of translator students and a survey are planned, with one experiment and one control group. The first step of the study is the implementation of the students translation tasks in class. A translation task was given to 20 students participating in the translation task. This second step was transferring the student translations into the computer and compiling a small scale student translation corpus (LPCorp). The third step was to code the students translations according to the typology of errors. The first three steps of the study have been completed. This is an ongoing project supported TUBITAK (The Scientific and Technological Research Council of Turkey) 2209-A - Research Project Support Programme for Undergraduate Students.

In the next phase of the study a new group of students will participate in the study to prevent participant bias. The students who are going to carry out the revision task will not be the same students who made

the translations. The experimental group of participating students will be trained to use the BNC corpus as a reference corpus for translation revision. After the training, the students will be given a revision task including erroneous sentences selected from the student translation corpus. The students will perform the revision tasks in a computer lab by consulting the BNC. On the other hand, the control group will refer to dictionaries or grammar reference books to make their revisions. By comparing the translations edited by the students, comparisons will be made between the classical editing method and the corpus consultation method, and a survey will be conducted to determine the students' attitudes towards different editing methods. The dependent variables in the study are the number of errors in different categories in the translations, the rate of correct correction of the errors, and the independent variables are the translation correction practices applied to both groups. T-test and Chi-square methods will be used as statistical methods.

Results

Since this is an ongoing project, the results of the study are yet to be completed. The results of the study are expected to contribute to translation training by experimenting with the use of corpus consultation in translation revision. We believe that corpus consultation has a potential in to be used in translation revision and it has an added benefit of increasing the student translators' autonomy and self-revision abilities.

Educating qualified translators who are equipped with 21st century skills, who can use technology well will provide more effective knowledge transfer in every field. In this way, faster progress can be made in the integration of our country with the world. In this sense, good translation also has an economic and commercial added value

References

Bennett, G. (2010). An Introduction to corpus Linguistics. Part 1 Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. https://www.press.umich.edu/pdf/9780472033850-part1.pdf

Biber, D., Conrad, S., Reppen, R. 1998. Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Biber, D., Reppen, R., & Conrad, S. (2007). Corpus linguistics : investigating language structure and use (pp. 12, 13). Cambridge Univ. Press.

Bulut, A.K. (2015). Translation Error. Istanbul: Translation Studies Publications

Kennedy, G. (1998). An Introduction to Corpus Linguistics. London: LONGMAN

Munday, J. (2016). Introducing translation studies: Theories and applications. Routledge.

Saint-Exupéry, A. (1943). Le petit prince (2015), Küçük Prens (translators. Cemal Süreya and Tomris Uyar), İstanbul: Can Art Publications Ltd.

Le sens d'un mot en FLE à travers le corpus de texte.

Agnieszka Dryjańska
Université de Varsovie
a.dryjanska@uw.edu.pl

Introduction

Le sens étant une notion clé de la linguistique cognitive, Wierzbicka soulignait que « to study the language without reference to meaning is like studying road signs from the point of view of their physical properties […] » (1996 : 3). Cela est également applicable à l'enseignement / apprentissage d'une langue étrangère. Or, bien que l'on connaisse depuis longtemps la nécessité d'apprendre les mots dans leur cotexte (Sardier, 2018), on ne cesse d'introduire de nouveaux lexiques sous forme de listes de mots dépourvus de leurs combinatoires lexicales, voire de leur sens, ce qui est observable dans des méthodes de français, p.ex. Edito B1. Afin de pallier ce problème, nous proposons de recourir à des corpus de textes qui permettent d'inférer le sens des mots à partir de leurs collocations (Zufferey, 2020). Ainsi, notre hypothèse est que l'emploi de corpus de textes permet d'aider à développer les compétences lexicale et sémantique englobant la connotation et les relations inter-lexicales telles que la collocation et l'équivalence en traduction, mais également des composantes des compétences générales savoir-apprendre et savoir-être, selon le Cadre européen commun de référence pour les langues (CECR) (2001).

Dans le cadre théorique de cette étude, nous allons éclaircir des notions qui sont au carrefour de l'approche cognitive et de l'approche fondée sur l'analyse de corpus comme la collocation, la combinatoire lexicale et le sens d'un mot. Nous réfléchirons ensuite sur la vision didactique du sens et sa place dans les compétences lexicale et sémantique telles qu'elles sont présentées dans le CECR. La partie empirique contient trois recherches-actions réalisées auprès d'étudiants en Philologie romane de l'Université de Varsovie. Dans chaque cas, les étudiants exploraient eux-mêmes des corpus de texte en français et en polonais afin d'observer des régularités combinatoires pour « faire des hypothèses sur le sens d'un mot » (Grossmann, 2011) et ses caractéristiques comme la prosodie sémantique ou la fréquence.

Cadre théorique

La formalisation du sens dans la linguistique de corpus basée sur les *lexicogrammatical patterns* s'explique par le fait que l'on peut identifier le processus cognitif qui conduit à la *conceptualisation* en analysant les structures linguistiques (Rewiś-Łętkowska, 2015). La notion lexique-grammaire qui découle de cette conception, forgée par M. Gross (« nous ne séparons pas cette dernière [grammaire] du lexique » (1981 : 48)), ou l'alliance lexique-syntaxe (Picoche, 1993 : 29 dans Nisubire, 2003 : 21) ont aussi des conséquences pour la Didactique des langues et cultures (DLC). Les éléments lexicaux constitutifs de la compétence lexicale énumérés dans le CECR ne se limitent pas aux mots isolés, mais englobent des expressions et locutions toutes faites, qui sont facilement accessibles grâce aux corpus de textes. Qui plus est, l'analyse de corpus peut s'avérer utile dans l'identification de la prosodie sémantique. D'après notre expérience, c'est une problématique très rare, voire inexistante dans les méthodes de français, et pourtant, elle peut être une source d'erreur, p. ex. le verbe *causer* en français s'associe le plus souvent à des collocatifs à connotation négative, ce qui n'est pas une règle en polonais.

De plus, vu une approche *contextuelle* au sens (Firth, 1957), l'acquisition du sens d'un mot étranger ne doit plus être le résultat de l'apriorisme lexicographique, mais peut émerger aux yeux de l'apprenant-chercheur à travers la complexité de la combinatoire lexicale et de nombreux contextes. Cela s'inscrit dans le *data-driven learning* favorisant une démarche heuristique en classe avec l'apprenant devenant « a language detective » (Johns, 1991).

Analysés dans la perspective contrastive entre la langue maternelle et la langue cible, certains aspects sémantiques s'avèrent même plus saillants et leur appréhension plus facile pour l'apprenant : la non-coïncidence collocationnelle rime avec la non-coïncidence sémantique des items lexicaux dans différentes langues et cultures (Dryjańska, Kazlauskienė, 2022).

Corpus et méthodologie

Corpus

Dans cette étude, nous recourons à la Leibniz Corpora Collection (LCC) (French et Polish) et Frantext en tant que corpus facultatif. Le premier offre un accès à des corpus contenant des données d'Internet en différentes langues, français et polonais inclus, le deuxième est un corpus français principalement littéraire.

Les corpus français dans la LCC utilisés dans le projet – French Mixed 2012, contenant 1,468,766,604 de mots et fra_news_2011_3M comprenant 63,125,248 de mots.

Le corpus polonais dans la LCC utilisé dans le projet – pol_newscrawl_2011 contient 96,476,260 de mots.

L'emploi didactique de la LLC s'explique par la simplicité de son interface, et notamment ses outils graphiques facilitant l'accès pour des utilisateurs non expérimentés et une comparabilité relativement fiable des résultats en français et en polonais, vu l'homogénéité de la méthodologie d'exploration dans tous les corpus de la collection. De plus, la LCC permet une recherche synchronique, car elle est fondée sur des ressources Internet récentes. Frantext offre des fonctionnalités plus complexes, voire plus difficiles pour les étudiants. Il permet de mener une recherche diachronique n'étant pas indispensable dans nos projets, mais pouvant la compléter.

Méthodologie

Les trois recherches-actions (méthodologies « entre théorie et pratique » (Richter, 2011)) conçues pour notre étude sont basées sur des corpus de texte avec la modalité *hands-on.*

Le premier projet, réalisé en première année (60 étudiants), était centré sur le processus de féminisation des noms de métiers et avait pour objet une analyse de contenus dictionnairiques et de données de corpus en français et en **polonais**. Le deuxième projet, proposé en troisième année (20 étudiants), était focalisé sur le lexique du champ sémantique propre au tourisme. Le troisième, conçu pour la première année de master (20 étudiants), visait une recherche lexicale dans le domaine social. Dans chaque cas, la tâche consistait à sélectionner des lexèmes à l'analyse, à les rechercher dans la LCC (et Frantext). Ensuite, il fallait « découvrir » leur sens à partir des données dans le corpus, conformément à la conception de Firth, sélectionner des exemples intéressants et les interpréter soit en se basant sur son propre savoir concernant notamment la langue polonaise ou rechercher une explication.

Résultats

L'analyse des résultats de tous les projets a fourni de nombreux exemples[39] qui ont confirmé que le corpus de texte était un outil possédant un fort potentiel didactique exploitable à différents niveaux de maîtrise de la langue française et permettant de développer différentes compétences, notamment lexicale et sémantique, mais également des compétences générales, comme savoir-apprendre et savoir-être.

Le développement de la compétence lexicale et sémantique à base de la combinatoire lexicale et la fréquence.

Tout d'abord, l'analyse de corpus a fourni, d'un côté, des collocations communes en polonais et en français, et de l'autre, des collocations spécifiques pour une des deux langues. Cela a donc permis d'identifier des similitudes et des ressemblances lexicales dans les deux langues. Tout d'abord, les étudiants ont identifié que les champs sémantiques de certains lexèmes « équivalents » sont plus larges dans une langue que dans l'autre. L'analyse des mots *hôtel* et *paysage* a fourni des collocations spécifiques au français, soit n'existant pas en polonais comme « hôtel particulier », « hôtel de ville », soit extrêmement rares en polonais comme « paysage audiovisuel » et « paysage médiatique. De plus, les étudiants ont identifié des collocations communes en polonais et en français – « krajobraz kulturowy » (fr. paysage culturel) et « krajobraz polityczny » (fr. paysage politique), étant intéressantes même pour le locuteur natif du polonais, car elles sont rarement utilisées dans la communication quotidienne.

L'étude a permis à quelques étudiants d'observer une ressemblance sémantique réalisée par différents moyens conceptuels et lexicaux. Cela peut être montré par les collocations qui existent uniquement soit en polonais soit en français comme « gare routière » (lit. dworzec drogowy), « dworzec autobusowy », (lit. gare de bus). Les étudiants ont également remarqué, ce qui est décrit par González Rey (2015 : 8), que certains lexiques avaient une biographie riche en histoire comme « dworzec PKS ». De plus, il y a des collocations qui existent dans les deux langues comme « violence verbale » (lit. przemoc werbalna) et przemoc psychiczna (lit. violence psychique), mais leurs fréquences diffèrent. En polonais, on parle plus fréquemment de « przemoc psychiczna » que de « przemoc werbalna ».

Qui plus est, les étudiants ont identifié une polysémie de certains lexèmes comme *mineuse* ou *maçonne*, ne figurant pas souvent dans les manuels scolaires, ce qui, d'un côté, enrichit les connaissances lexicales des étudiants et, de l'autre, met en lumière l'un des arguments avancés contre la féminisation des noms de métiers. En Pologne et en France, on critique l'emploi de certaines formes car elles ont déjà un référent, et, qui plus est, c'est souvent un référent non humain, ce qui s'avère problématique pour les usagers d'une langue (Lenoble-Pinson, 2008, Szpyra-Kozłowska, 2019).

Un autre résultat de l'étude est l'observation de la prosodie sémantique. Des étudiants ont observé que les collocations d'un lexème peuvent être catégorisées comme celles ayant une connotation positive ou une connotation négative. Dans le cas du lexème *séjour*, l'on distingue deux types de collocations en polonais et en français : « séjour agréable », « przyjemny pobyt », « séjour de vacances », mais également « pobyt nielegalny » « séjour en prison », etc. Cependant, il y des mots qui ont une prosodie sémantique dominante comme *braterstwo* (fr. fraternité) apparaissant plus fréquemment en polonais avec des collocatifs à connotation négative comme *wojna* (fr. guerre), *broń* (fr. arme), *oszust* (fr. tricheur), contrairement aux résultats pour son équivalent français révélant une prosodie sémantique positive.

Savoir-apprendre : aptitudes heuristiques

Ensuite, la comparaison entre les définitions dictionnairiques et le résultat de la « fouille » dans la LCC a révélé que certaines formes féminines figurant dans les dictionnaires étaient très rares, leurs classes de fréquence étant hautes dans French Mixed 2012 comme avocate (14), vendeuse (15) ou factrice (19). Cela montre aussi que le corpus de texte peut jouer un rôle complémentaire en enrichissant l'information dictionnairique en DLC.

[39] Pour les contraintes rédactionnelles, tous les exemples ne peuvent pas être cités ici.

Il s'est également avéré que les étudiants étaient capables de maîtriser rapidement les fonctionnalités de la LCC, et même celles de Frantext. Dans le cas de ce deuxième, il y a des étudiants qui ont trouvé eux-mêmes la fonctionnalité qui leur a permis d'explorer uniquement des textes du XX^{ème} siècle. Dans la LLC, afin de comparer la fréquence des lexèmes, ils ont recouru à la classe de fréquence, vu que les tailles des corpus polonais et français étaient différentes donc la comparaison des fréquences absolues n'était pas possible. Ensuite, des limites de l'analyse quantitative dans la LCC, p. ex. quant aux formes épicènes ont été remarquées. Dans ce cas, l'identification d'un référent féminin exige une analyse « manuelle » de la concordance.

Savoir-être

Un autre avantage didactique des projets, inattendu, est l'intérêt pour les problématiques étudiées éveillé, semble-t-il, grâce à l'autonomie et à la dimension heuristique des activités proposées. À notre sens, cela se confirme par l'emploi fréquent des mots *ciekawy*, *intéressant*, warto, *valoir* dans les rapports étudiants, mais également par des recherches facultatives soit sur le corpus Frantext soit dans la presse et dans les documents officiels de l'Académie française traitant la problématique de la féminisation des noms de métier.

Nous avons observé que l'apprenant s'activait face au corpus de texte, informateur riche et puissant, mais « passif », et que, par sa passivité, ce dernier rendait l'apprenant actif. Les activités basées sur de nombreux extraits de textes et des combinaisons lexicales permettent un apprentissage lexical inductif par excellence suivant le modèle « I(dentify) – C(lassify) – G(eneralise) » de Johns (1991).

Références bibliographiques

Cadre européen commun de référence pour les langues – Apprendre, Enseigner, Évalur (2001). Conseil de l'Europe.

Dryjańska, A., Kazlauskienė, V. (2022). Le sens de *fête* en polonais, en lituanien, en français et sa (non)coïncidence collocationnelle. *Annales Universitatis Paedagogicae Cracoviensis. Studia Linguistica*, 17, 20-42.

Dufour, M. et al. (2018). *Édito B1. Méthode de français*. Paris : Les Éditions Didier.

Dubois, J. et.al. (1973). *Dictionnaire de linguistique*. Paris : Librairie Larousse.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. Dans J. R. Firth (dirs.), *Studies in Linguistic Analysis*, l-32. Basil Blackwell Oxford. Consulté le 05.02.2023 http://cs.brown.edu/courses/csci2952d/readings/lecture1-firth.pdf

González Rey, I. (2015). La didactique du français idiomatique. EME Éditions.

Gross, M. (1981). Les bases empiriques de la notion de prédicat sémantique. *Langages*, 63, 7-52. Consulté le 05.02.2023 https://www.persee.fr/doc/lgge_0458-726x_1981_num_15_63_1875

Grossmann, F. (2011). Didactique du lexique : état des lieux et nouvelles orientations. *Pratiques : linguistique, littérature, didactique*, 149-150, 163-183. Consulté le 05.02.2023 Didactique du lexique : état des lieux et nouvelles orientations (hal.science)

Lenoble-Pinson, M. (2008). Mettre au féminin les noms de métier : résistances culturelles et sociolinguistiques." *Le fraçais aujourd'hui*, 163/4, 73-79.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. Dans T. Johns & P. King (dirs.), *Classroom Concordancing. ELR Journal, 4,* 1-16. Consulté le 05.02.2023 https://lexically.net/wordsmith/corpus_linguistics_links/Tim%20Johns%20and%20DDL.pdf

Kosecki, K. (2015). Ethnic and gender stereotypes in signed languages : A cognitive linguistic view. Dans M. Kuźniak M. & A. Libura A. & M. Szawerna (dirs.), *From Conceptual Metaphor Theory to Cognitive Ethnolinguistics,* 38-50. Peter Lang.

Nisubire, P. (2003). Développer la compétence lexicale en classe de français langue seconde. *La Lettre de l'AIRDF,* 33, 20-26. Consulté le 05.02.2023 Développer la compétence lexicale en classe de français langue seconde - Persée (persee.fr)

Rewiś-Łętkowska, A. (2015). Conceptualization of fear in English and Polish. Dans M. Kuźniak & A. Libura & M. Szawerna (dirs.), *From Conceptual Metaphor Theory to Cognitive Ethnolinguistics,* 79-91. Peter Lang.

Richer, J.-J. (2011). Recherche-action et didactique du FLE. *Synergies Chine*, 6, 47-5. Consulté le 05.02.2023 8https://gerflint.fr/Base/Chine6/richer.pdf

Sardier, A. (2018). Inciter les élèves à recourir au cotexte pour inférer le sens des unités lexicales. *La Lettre de l'AIRDF*, 64, 36-42. Consulté le 05.02.2023 Inciter les élèves à recourir au cotexte pour inférer le sens des unités lexicales - Persée (persee.fr)

Szpyra-Kozłowska, J. (2019). Premiera, premierka czy pani premier? Nowe feminatywy w ujęciu ankietowym. *Język Polski*, 2, 24-40.

Warren, M. (2011). Using Corpora in the learning and teaching of phraseological variation. Dans A. Frankenberg-Garcia & L. Flowerdew & G. Aston (dirs.), *New trends in corpora and language learning,*153-166. Continuum.

Wierzbicka, A. (1996). *Semantics. Primes and universals*. New York: Oxford University Press.

Zufferey, S. (2020). *Introduction à la linguistique de corpus*. London : Iste éditions.

Explorations textométriques d'un corpus foucaldien. Le Désordre *des familles* : au plus près de la naissance du genre du rapport

Hugo Dumoulin[1]
[1] Laboratoires Modyco & Sophiapol, Université Paris Nanterre
hugo.dumoulin@parisnanterre.fr

Introduction

En 1968, lorsque le cercle des étudiants en épistémologie de l'ENS l'interroge sur sa méthode, Foucault répond : « À tant d'incertitudes, je voudrais substituer l'*analyse du discours* lui-même dans ses conditions de formation, dans la série de ses modifications et dans le jeu de ses dépendances et de ses corrélations » (1994, I : 682, nous soulignons). Si une telle déclaration ne manque pas de situer le philosophe parmi les penseurs fondateurs de l'analyse du discours, elle n'en est pas moins énigmatique : comment repérer séries, dépendances, corrélations, conditions de formation du discours ? Y'a-t-il une théorie qui formaliserait les structures de la discursivité ? Déclarant finalement que « le temps n'est pas encore venu de la théorie » (1969) Foucault ne nous a pas laissé de clef disciplinaire définitive pour contrôler ses procédures de découverte. Il nous laisse cependant des corpus, constitués par « la mise en travail de problématiques philosophiques au sein d'archives historiques » (Revel 2010).

Cette étude veut montrer à partir d'un exemple qu'une lecture de linguistique de corpus de ces matériaux est possible, et que les explorations textométriques permettent de répondre à certaines questions laissées ouvertes par Foucault. Il se trouve que Foucault a publié quelques-uns de ses corpus, nous rendant la tâche plus facile : c'est le cas du *Désordre des familles* (Foucault & Farge 1982), abordé sous l'angle d'une lecture outillée dans notre thèse, de laquelle cette étude est extraite.

Corpus et méthodologie

Corpus

Comme l'indiquent les éditeurs des *Dits et Écrits* (1994, III : 237), Foucault a longtemps poursuivi le projet d'une grande publication des archives de l'enfermement à l'hôpital général et à la Bastille, sur lesquelles il n'a cessé de travailler depuis l'*Histoire de la folie* (1961). De ce grand projet cependant, il ne reste que *Le Désordre des familles*, publié en collaboration avec Arlette Farge dans la collection corpus de Gallimard (1982), et présentant 287 textes tirés des dossiers de police versés aux archives de la Bastille pour les années 1728 et 1758 conservées à la bibliothèque de l'Arsenal à l'époque de leur consultation par Foucault et Farge.

Ces dossiers sont liés à l'institution des lettres de cachet, ordres émis par le lieutenant général de police par délégation du roi à dessein d'enfermer un individu à la Bastille sans jugement. En rupture avec l'historiographie révolutionnaire, Foucault et Farge montrent que ces lettres sont émises massivement suite à la demande des familles elles-mêmes (1982 : 9) pour régler tout un ensemble d'affaires d'ordre privé qui vont du conflit de voisinage à des questions plus graves qui relèveraient aujourd'hui du pénal. Dans les dossiers l'on trouve ainsi des lettres écrites (via l'écrivain public) par les familles requérant l'enfermement d'un tiers (les « placets »), des lettres écrites par un ensemble de personnages se présentant comme les garants des précédentes, enfin des textes écrits par les policiers faisant la synthèse de leurs démarches d'enquête suite aux demandes et rendant un avis sur l'ordre d'enfermement.

Le livre donne accès au corpus d'archives à travers plusieurs couches d'un traitement accompli par Foucault et Farge. Les deux historiens transcrivent et dactylographient les manuscrits d'archives, tout en indiquant leurs difficultés de lecture de certains mots. Ils normalisent aussi une mise en page d'origine où du texte se trouve parfois écrit dans la marge ou en surimpression, ce dont ils gardent systématiquement la trace en le présentant de manière homogène en crochets dans le fil du texte ou en annexe à la fin de celui-ci. Surtout, Foucault et Farge ajoutent à la classification archivistique certains éléments de classement qui sont le reflet de leur propre lecture et qui se concrétisent dans le découpage du livre en parties thématiques. Du reste, cette lecture historienne se fait discrète dans le livre : la part du commentaire est grandement réduite par rapport à l'espace alloué à la retranscription de pièces[40]. Pourtant, elle n'est pas absente, et c'est en gardant trace de ces multiples couches de lecture par l'annotation que l'on peut espérer capter quelque chose des procédures de découverte des deux historiens à travers par la relecture textométrique.

Méthodologie

Publié tardivement par rapport à la période du travail entre Foucault et Farge, *Le Désordre des familles* est marqué par la problématisation foucaldienne du milieu des années 1970. À ce moment, Foucault réalise une histoire de ce que l'on pourrait appeler les « technologies de pouvoir » (Foucault 1975), qui met au jour l'émergence à partir du XVIII^e^ siècle du régime des « disciplines » (Foucault 1973), clef de voute des deux infinitifs du titre foucaldien : « surveiller » et « punir » (Foucault 1975), lesquels articulent « savoir » et « pouvoir » (Foucault 1973). Foucault et Farge avancent l'hypothèse que l'institution des lettres de cachet est une des premières mises en œuvre de la technologie de pouvoir « disciplinaire » (Foucault & Farge 1982 : 346). Les lettres de cachet seraient donc l'un des premiers instruments de « pouvoir-savoir ». Notre hypothèse est qu'une enquête de linguistique de corpus sous la forme d'une analyse des données textuelles permet d'éprouver l'hypothèse de Foucault et Farge, en cernant le fonctionnement *discursif* de ce dispositif de normalisation disciplinaire.

On avance que l'insertion de ces textes dans un certain genre joue un rôle décisif. Foucault lui-même indique que les « disciplines » s'attachent à certains types d'écrits administratifs qui émergent à la fin du XVIII^e^ siècle, Foucault parlant du début de « l'ère du *rapport* comme forme des relations entre savoir et pouvoir » (1973 : 238 nous soulignons). Si Foucault ne cite pas Bakhtine, il est manifeste qu'il s'intéresse à ce que le dernier définit comme des *genres* de discours situés au carrefour du langagier et du social et référés à certaines sphères d'activité (Bakhtine 1984 : 269). Et en effet, bien qu'empruntant des formes diverses, les textes du corpus gardent pour fonction de rapporter et de rendre compte, ce qui autorise de les aborder à partir de catégorie du macro-genre du « rapport », aux contours bien définis : double modalité de la « description » des faits et de leur « évaluation » axiologique (en vue de la prescription d'une action), caractère « adressé » de textes marqués par le différentiel hiérarchique entre un scripteur et le commanditaire auquel il s'adresse (Née, Oger, Sitri, 2017).

Notre méthode d'acquisition repose sur la conversion du corpus textuel d'archives transcrites et imprimées dans le livre du *Désordre des familles* en base de données textuelles (.xml) exploitable par des logiciels de textométrie. Partant d'une version .txt du livre, l'on réalise des choix de balisage visant à conserver certaines caractéristiques des textes en métadonnées ainsi que certains aspects de la composition textuelle à travers des balises (Figure 1). Parmi les métadonnées de chaque texte l'on conserve des informations afférentes aux différentes couches d'exploitation du corpus : certaines représentent la classification adoptée à la bibliothèque de l'Arsenal, d'autres sont issues du traitement par Foucault et Farge du corpus. Parmi les premières, on trouve la date d'écriture de la pièce (« 1728 »), l'identifiant de celle-ci (« 11009f95 »), ainsi que le dossier dans lequel elle se trouve (représenté par nom d'affaire : « Barbe Blondel Duponchel »). Parmi les secondes, l'on trouvera le rang de la pièce dans l'ordre de présentation de chaque dossier dans le livre, ainsi que le regroupement des dossiers en thèmes selon les deux niveaux de classement du livre : une première division en deux grandes catégories « La discorde des ménages (DM) » et « Parents et Enfants (PE) », une deuxième division définissant dans

[40] Sur les 365 pages du livre, seules 55 sont consacrées à la présentation et aux analyses du corpus.

chaque catégorie des sous-thèmes comme « le déshonneur de l'errance » ; cette annotation donne en sortie une valeur d'attribut complexe du type « PE_le déshonneur de l'errance » (Figure 1).

Pour représenter certains aspects de la composition textuelle des pièces du corpus à travers le balisage, l'on retiendra des balises classiques représentant en-tête (<head>), division en paragraphes (<p>), et signatures (<signed>) de chaque pièce, ainsi qu'une balise servant à représenter le fait que, sur une même pièce, plusieurs scripteurs peuvent parfois cohabiter : comme l'indiquent Foucault et Farge, la lettre du garant ou le rapport de police peuvent être écrits parfois à même le placet d'origine, dans la marge ou au verso de celui-ci. Ces segments de texte, bien que faisant partie de la même pièce d'archive, ont des statuts différents : ainsi, à travers les différentes valeurs d'un attribut « type » de la balise l'on gardera trace non pas seulement des différents scripteurs, mais aussi des différentes démarches dans lesquelles ils insèrent explicitement leur écrit. Ce balisage est facilité par l'abondance des énoncés performatifs dans ces textes – qui représentent l'acte que le sujet dit accomplir par son dire (Authier-Revuz 2020 : 27) – au sein de formules dont on peut mesurer par ailleurs le caractère stéréotypé à l'aide d'un calcul des segments répétés[41]. L'on encode dix types de démarches par dix valeurs de l'attribut type[42].

Après acquisition comme base de données textuelles, le corpus DÉSORDRE regroupe 77 049 occurrences pour 287 textes, soit une moyenne de 266 mots par texte.

Résultats

L'on aborde dans cette étude deux grandes analyses contrastives réalisées sur la base du balisage précédent.

L'ancrage lexical des contrastes dans la représentation de la norme

A l'aide du logiciel TXM (Heiden et *al*. 2010) on lemmatise le corpus et on réalise un balisage syntaxique de celui-ci selon l'étiquetage TreeTagger. L'on peut alors réaliser une analyse factorielle des correspondances (Benzécri 1973) entre les formes graphiques présentes dans le corpus et une partition de celui-ci selon les thèmes retenus par Foucault et Farge (Figure 2). Le résultat obtenu est particulièrement éclairant : le premier axe de l'AFC se laisse interpréter comme une opposition est-ouest entre parentalité d'une part et conjugalité, selon laquelle s'accordent exactement les parties du corpus inscrites dans les macro-thèmes de Foucault et Farge « Parents-enfants » (PE sur le schéma) – à l'est –, et « Discorde des ménages » (DM) – à l'ouest. En effet, les formes les plus contributives à l'axe 1 se trouvent partagées entre l'est – « fils » (contribution de 7,5%), « père » (5%) », « mère » (2%) –, et l'ouest – « femme » (11,5%) et « mari » (6%). À l'ouest se trouvent les formes employées pour désigner les époux à l'intérieur du ménage, à l'est pour désigner les membres de la relation parentale.

Le second axe se laisse interpréter comme une opposition nord-sud entre plaintes visant des hommes et plaintes visant des femmes : les formes les plus contributives à l'axe 2 se trouvent partagées entre formes désignant des hommes au nord – « il » (9,5%), « fils » (3,5%), « ledit » (1%) –, et des femmes au sud – « fille » (6%), « elle » (4%), « ladite » (3,5%). Cet axe permet de contraster selon une opposition homme/femme le découpage thématique proposé par Foucault et Farge, ce que confirme d'ailleurs la

[41] Ainsi par exemple un segment relativement long comme « X représente très humblement à votre Grandeur le / que » affiche une fréquence de 69 occurrences dans le corpus selon un calcul réalisé par Le Trameur (Fleury & Zimina 2014).

[42] La liste des valeurs est : "enfermement" = scripteur demandant l'enfermement d'un tiers ; "libération" = scripteur demandant la libération d'un tiers ; "libération_soi" = scripteur demandant sa propre libération ; "libération_repentir" = scripteur demandant la libération d'un tiers dont il a précédemment demandé l'enfermement ; "enfermement_garant" = scripteur se présentant comme un garant pour enfermer un tiers ; "libération_garant" = scripteur se présentant comme un garant pour libérer un tiers ; "police_rapport"= scripteur policier rédigeant un rapport d'enquête ; "police_DR_enfermement" = scripteur policier rapportant le texte d'un placet : "police_DR_libération" = scripteur policier rapportant le texte d'une demande de libération ; "correspondance" = lettre de correspondance versée au dossier.

position du thème de « la débauche des *maris* » au nord et de celui de « l'enfermement des *épouses* » au sud.

Une fois ce cadre interprétatif posé, l'on peut se pencher sur la position de différents lexèmes contribuant à la représentation de l'anormalité elle-même, et observer ainsi leur contribution aux contrastes du corpus. Ainsi, l'on observe sur l'axe 1 que les formes « argent » et « livres » sont significativement attirées à l'est, c'est-à-dire vers le rapport parent/enfant, tandis que les formes « conduite » et « vie » sont significativement attirés à l'ouest, du côté du rapport entre époux. Les parents reprochent à leurs enfants leur prodigalité matérielle, tandis que les époux se reprochent mutuellement leur mauvaise conduite morale. Le long de l'axe 2, l'on est attentif à la présence des formes « jours » et « enfants » au nord – du côté des plaintes des femmes visant les hommes –, par opposition à la présence de « débauche » et « libertinage » au sud – du côté des plaintes des hommes visant des femmes. Des retours au texte permettent d'éclairer cette répartition : alors que les femmes reprochent aux hommes leur dissipation (« il la consomme tous les *jours* en dépenses ») et leur absence dans la gestion des affaires du foyer (concernant les *enfants* en particulier), les hommes se plaignent de l'inconduite supposée de leurs concubines. La norme n'est donc pas la même pour tout le monde : figure bariolée, le scandale de l'anormal change de forme en fonction des situations où il est invoqué.

L'analyse factorielle confirme l'ancrage *lexical* du découpage par thèmes de la « mauvaise conduite » selon Foucault et Farge. Cette détermination lexicale de l'expression de l'anormalité met en avant deux choses : d'une part il y a un ancrage discursif du dispositif de normalisation disciplinaire : ce sont certains lexèmes (« argent », « conduite », « enfants », « libertinage ») qui contribuent à fixer la représentation de l'anormal dans les différentes situations où il surgit ; d'autre part et conséquemment, l'on peut arguer que la procédure de classification par Foucault et Farge des thématiques du corpus – et donc des aspects de la norme – repose plus ou moins inconsciemment sur le repérage de ces précédents lexèmes dans les textes, dans la mesure où notre analyse statistique met en valeur leur caractère significatif pour la classification de Foucault et Farge.

L'ancrage énonciatif des contrastes entre positions socio-discursives

L'on réalise alors une deuxième analyse factorielle des correspondances entre les lemmes du corpus et une partition selon les types de scripteurs repérés par leurs démarches (Figure 3). La représentation obtenue est marquée la domination écrasante de l'axe principal (45% de l'inertie totale). Cet axe se laisse interpréter selon une opposition entre à l'est des pronoms personnels de 1ère et 2ème personne ainsi que certains termes d'adresse (« monsieur »), et à l'ouest des termes d'adresse comme « Monseigneur », ou des déterminants comme « Votre ». L'opposition radicale entre est et ouest est celle de la distance entre l'énonciateur et son interlocuteur : à l'est l'on a un rapport de relative égalité, à l'ouest il y a une distance hiérarchique écrasante. Ainsi, la répartition des types de termes d'adresse entre les scripteurs s'avère structurante des contrastes entre démarches au sein du corpus, ce qui confirme bien la prégnance sur ces textes de leur identité générique de rapports, c'est-à-dire de textes *adressés*.

L'on observe alors un phénomène spectaculaire : les textes des garants, qu'ils soient destinés à soutenir une demande d'enfermement ou de libération, se trouvent déportés à l'est ; à l'inverse, les textes des familles populaires, qu'ils visent l'enfermement ou la libération, sont déportés à l'ouest. Cela est décisif : les deux démarches (opposées d'un point de vue pragmatique) de la demande d'enfermement et de libération apparaissent tout de même moins différentes *énonciativement* que ce qu'on peut désigner comme les « positions socio-discursives » des scripteurs (petit peuple des « misérables », garants issus du clergé, police)[43].

Ces positions, hiérarchiquement articulées entre elles, soulignent le différentiel social qui fonctionne à l'arrière-plan du dispositif de normalisation disciplinaire. Cette répartition en trois « positions discursives » est confirmée par une classification ascendante hiérarchique des types de démarches selon

[43] L'on peut capture les origines sociales des types de scripteurs en s'intéressant aux signatures qui figurent au bas des pièces, lesquelles comportent presque systématiquement la profession des signataires.

les lemmes employés (Figure 4). La généricité de ces textes participe donc à imprimer discursivement un ensemble de positions sociales différenciées.

Conclusion

Ainsi, l'on conclut de ces explorations que les discours et leurs caractéristiques génériques participent à l'impression sur les sujets de normes de conduite et d'une hiérarchie sociale, précisant par-là le fonctionnement du « pouvoir-savoir » décrit par Foucault. Cette étude a une vocation méthodologique : en gardant trace de la lecture opérée par Foucault et Farge, l'on peut indirectement éclairer leurs procédures de découverte. Ainsi l'on montre que la classification thématique des dossiers du corpus – et, partant, des différents aspects que peut prendre la norme – s'attache à la présence statistiquement significative de certains lexèmes dans les textes. De même, le repérage d'un dispositif de hiérarchisation sociale dans l'institution des lettres de cachet est attaché à la répartition statistiquement significative de différentes formes d'adresse, qui contribue à contraster le corpus selon les démarches des scripteurs, qui forment autant de sous-genres (placet, lettre de garant, rapport de police).

Cette étude s'ouvre sur les autres analyses menées dans la thèse, qui nous conduisent, au-delà des réflexions foucaldiennes, à mettre en évidence l'influence de l'idéologie du petit artisanat parisien du XVIIIe siècle (Soboul 1981) sur l'état du genre du rapport à la même période. En menant sur Iramuteq (Marchand & Ratinaud 2015) une classification descendante hiérarchique « méthode Reinert » (1983) (Figure 5) on voit apparaître d'autres thématiques structurantes du corpus, comme par exemple la « matérialité économique » (classe 2). En menant des analyses déductives (calcul des spécificités, analyses des cooccurrences), on peut observer le comportement de formes de « préconstruit » caractéristique de l'emprise d'une idéologie sur le discours (Pêcheux 1975 : 152).

Figures

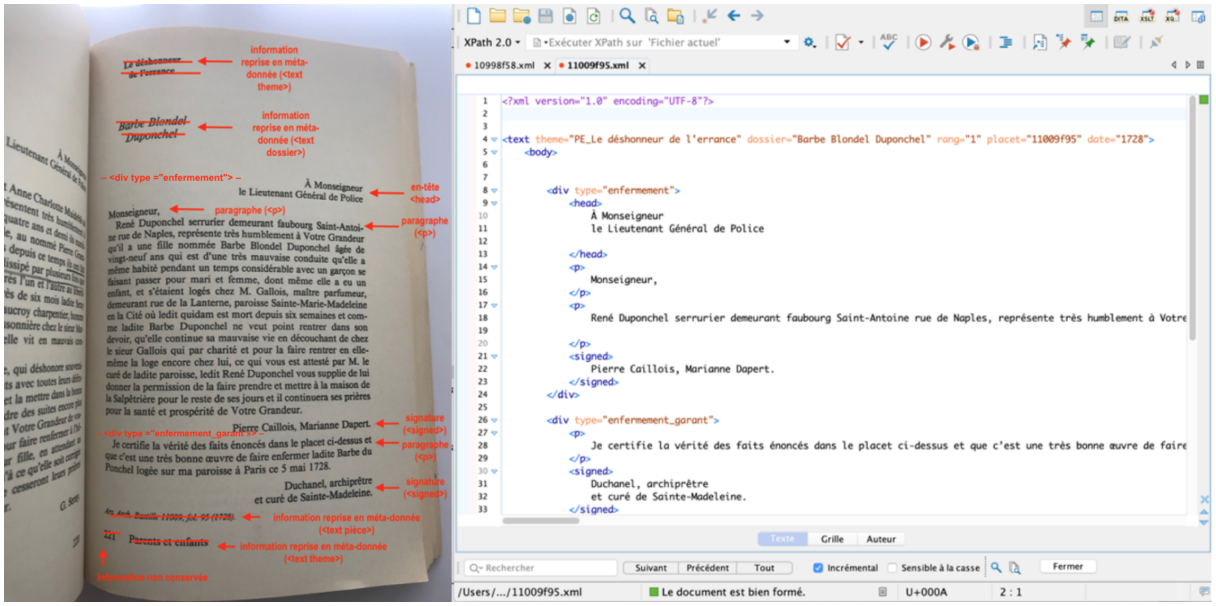


figure . 1 : Le balisage du corpus (à gauche une page de l'édition Gallimard, à droite la vue d'un fichier .xml dans OxygenXML)

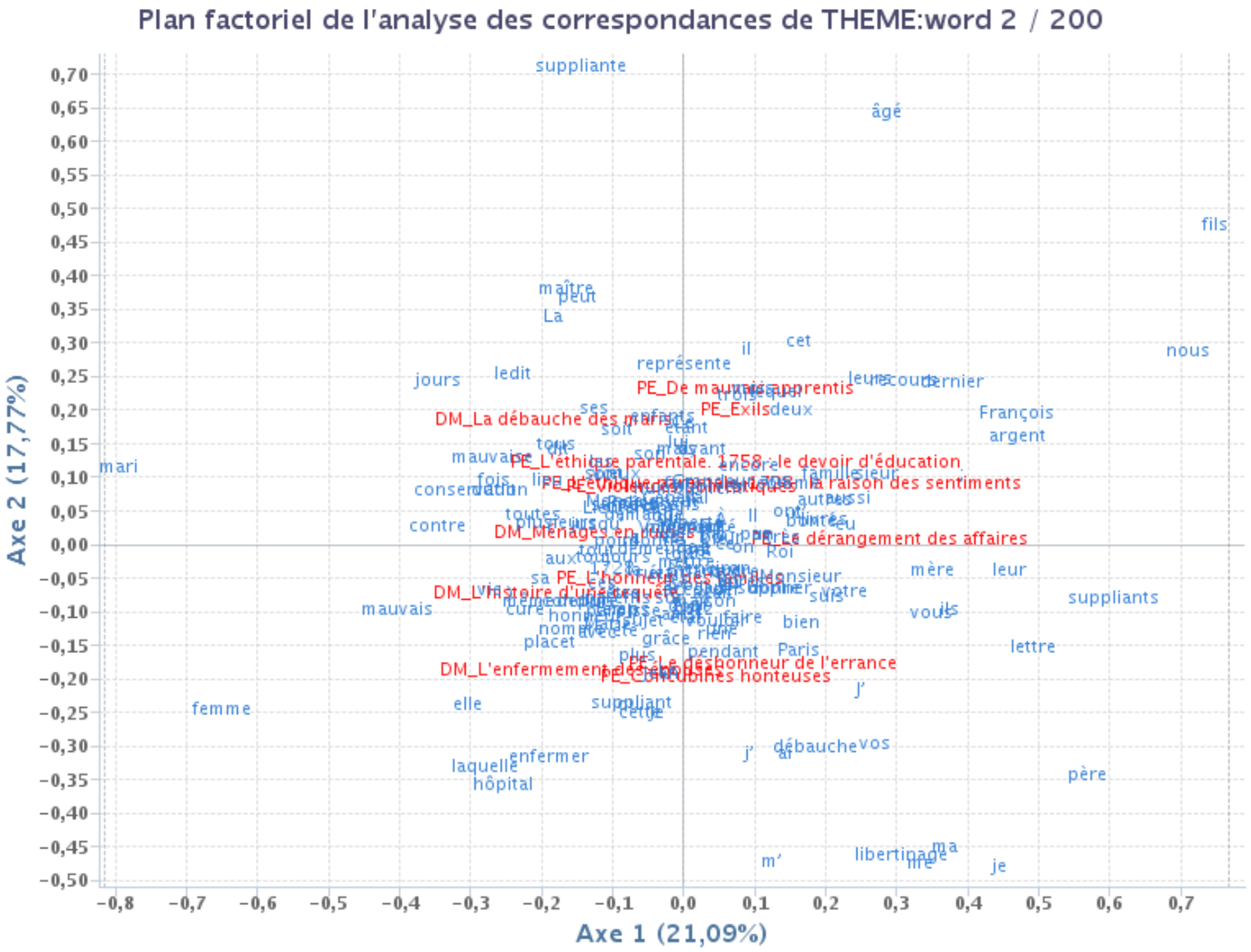


figure . 2 : AFC forme graphique × thème (fmin = 2, χ^2=5859,7)

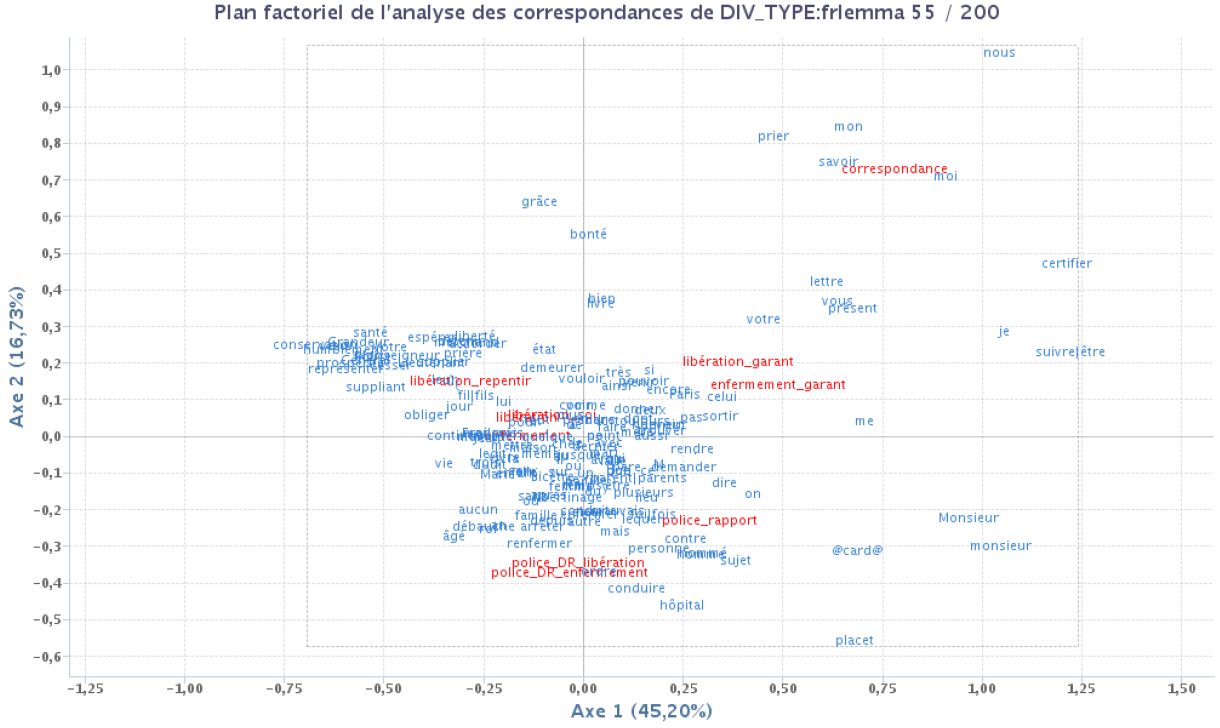


figure . 3 : AFC lemme × type de (fmin = 55, χ^2=7571,05)

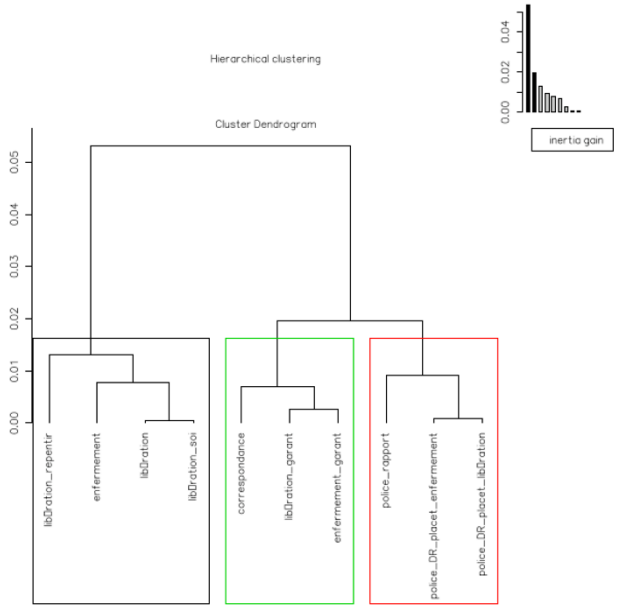


figure .4 : Classification ascendante hiérarchique du tableau de contingence types de × lemmes

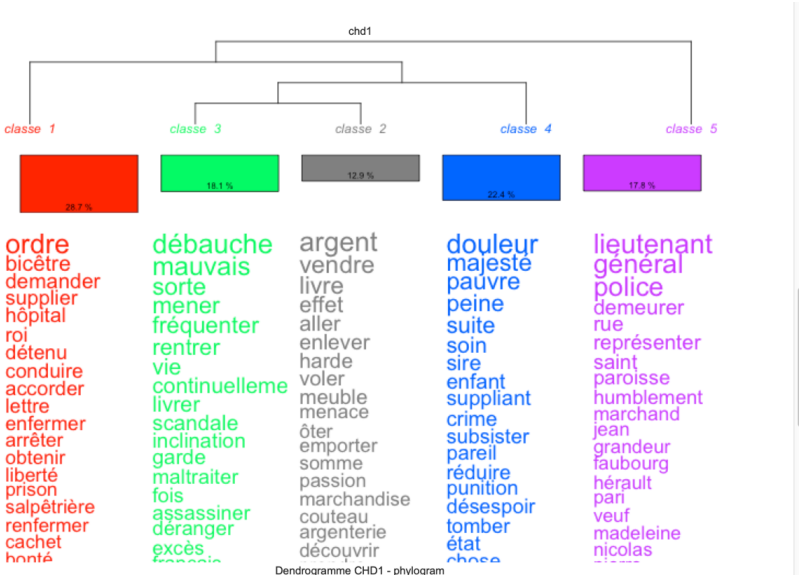


figure . 5 : Classification descendante hiérarchique méthode Reinert (Iramuteq)

Références bibliographiques

Authier-Revuz, J., 2020, *La Représentation du discours autre. Principes pour une description*, Paris, De Gruyter, 685 p.

Benzécri, J.-P., 1973, *L'Analyse des données*, Paris-Bruxelles-Montréal, Dunod, 2 vol., 1234 p.

Dumoulin, H., 2022, *Les théorisations du discours de Michel Pêcheux et Michel Foucault à la lumière du concept d'énonciation*, thèse de doctorat soutenue le 9 décembre 2022, Université Paris Nanterre, 700 p.

Foucault, M. & Farge, A., 1982, *Le Désordre des familles. Lettres de cachet des Archives de la Bastille*, Paris, Gallimard-Julliard, 362 p.

Foucault, M., 1961, *Histoire de la folie à l'âge classique*, Paris, Gallimard, 583 p.

Foucault, M., 1969, *L'Archéologie du savoir*, Paris, Gallimard, 275 p.

Foucault, M., 1973, *La société punitive*, Paris, Gallimard, Hautes Études, 349 p.

Foucault, M., 1975, *Surveiller et punir,* Paris, Gallimard, 318 p.

Foucault, M., 1994, *Dits et Écrits*, 4 volumes, Paris, Gallimard, 3427 p.

Fleury, S. & Zimina, M., “Trameur : A Framework for Annotated Text Corpora Exploration”, *Proceedings of COLING 2014, the 25th international Conference on Computational Linguistiques : System Demonstrations*, Dublin, pp. 57-61.
Heiden, S., Magué, J.-P., Pincemin, B., 2010, « TXM : Une plateforme logicielle open-source pour la textométrie-conception et développement », *JADT 2010 : actes des 10èmes Journées internationales de l’Analyse statistique des Données Textuelles*, Rome, vol. 2, n°3, pp. 1021-1031.
Marchand, P., Ratinaud, P., 2015, « Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l’Assemblée nationale (1998-2014) », *Mots. Les langages du politique*, 2015/2, n°108, pp. 57-77.
Née, E., Oger, Cl., Sitri, F., 2017, « Le rapport : opérativité d’un genre hétérogène », *Mots*, n°114.
Reinert, M., 1983, « Une méthode de classification descendante hiérarchique : application à l’analyse lexicale par contexte », *Les Cahiers de l’Analyse des Données*, 1983, vol. VIII, n°2, pp. 187-198.
Revel, J., 2010, *Foucault une pensée du discontinu*, Paris, Mille et une nuits, 333 p.
Soboul, A., 1981, *Comprendre la Révolution. Problèmes politiques de la Révolution française, 1789-1797*, Paris, Maspero, 379 p.

Étude diachronique comparative des adverbes *evidentemente* et *obviamente* dans la langue espagnole écrite : deux adverbes pour une même idée d'évidence ?

Catline Dzelebdzic [1]
[1] Laboratoire CeRLA, Université Lumière Lyon 2
c.dzelebdzic@univ-lyon2.fr

Introduction

Evidentemente et *obviamente* sont deux adverbes au sens particulièrement proches en espagnol actuel : les adjectifs dont ils sont issus sont ainsi définis comme « cierto, claro » ('certain, clair') pour *evidente* et « muy claro » ('très clair') pour *obvio* (Diccionario de la lengua española 2022). S'il existe des études diachronique et synchronique sur chacun d'eux individuellement (Sánchez Jiménez 2013 pour le premier aspect et Torner Castells 2016 pour le second) et un traitement lexicographique en espagnol synchronique (Fuentes Rodríguez 2009), une comparaison systématique n'a pas encore été réalisée. Or, une étude approfondie en diachronie récente, centrée sur le XXe siècle, période où émerge *obviamente*, servirait à connaître leurs différences de façon plus précise, dans la mesure où celui-ci commence à apparaître dans les textes au XIXe siècle alors *qu'evidentemente* y est utilisé depuis le XVe siècle. Une telle différence implique donc que les deux adverbes connaissent des processus d'évolution qui ne sont pas similaires.

Corpus et méthodologie

Corpus

Cette étude se centre autour de l'espagnol écrit : ce choix est fondé sur une caractéristique des adverbes en *-mente* espagnols, leur association particulière à la langue écrite et formelle (Hummel 2012 : 251-252 et 2018 : 122-124). Nous considérerons tout de même les occurrences extraites de documents oraux dans les corpus que nous utilisons, avec les précautions décrites plus bas.

Pour mener à bien cette étude, nous utilisons deux corpus espagnols, le *Corpus del Diccionario Histórico de la Lengua Española* (CDH) et, de façon complémentaire, le *Corpus de Referencia del Español Actual* (CREA). Tous deux sont des bases de données de grande taille (environ 38 millions de mots pour le premier dans sa version *Nuclear* et environ 200 millions pour le second), diachronique pour le CDH et synchronique pour le CREA. Ces deux corpus permettent de réaliser une analyse complète de la langue écrite, soit le cadre de cette étude, en ce qu'ils contiennent une grande diversité de documents écrits : littéraires, scientifiques, de didactique ou vulgarisation, des essais, des textes de presse, …. Par ailleurs, le CREA contient également des transcriptions d'extraits oraux, mais relevant souvent d'un cadre formel (télévision, radio, …), de sorte que nous pourrons uniquement proposer des hypothèses pour cette partie de l'espagnol oral.

Méthodologie

Les 874 occurrences de *obviamente* et les 3.014 occurrences de *evidentemente* ont été extraites directement dans les bases de données. Nous avons ensuite procédé à une analyse afin de préciser leurs propriétés selon des critères syntaxiques (position dans la phrase, présence de marques de ponctuation), sémantico-pragmatiques (fonction sémantique, effet de sens particuliers) et textuels (type de discours, type de texte).

Résultats

L'analyse des occurrences permettra, dans une perspective globale, de différencier l'évolution de *evidentemente* et de *obviamente* au XXe siècle et de retracer l'émergence du second adverbe au cours de ce même siècle : la comparaison mettra ainsi en évidence l'existence d'une analogie où l'adverbe le plus ancien, *evidentemente*, fonctionne comme un modèle. En effet, celui-ci connaît un processus évolutif de plusieurs siècles qui le mène à acquérir un emploi comme marqueur discursif, tandis cet emploi est présent dès les premières occurrences de *obviamente* dans le CDH. De la sorte, cette évolution ne peut pas être comprise sans considérer le parcours de *evidentemente* ainsi que son profil d'emploi dans la langue espagnole écrite au XIXe siècle, quand *obviamente* commence à être utilisé.

De plus, la prise en compte des différents critères issus du cotexte et du contexte d'emploi nous servira à identifier les variables qui s'associent le plus à l'emploi de chaque adverbe, afin d'en dégager leurs spécificités au-delà de la simple différence en termes de fréquence d'emploi dans le CDH et dans le CREA. Nous nous attacherons en particulier à vérifier dans les documents écrits et les transcriptions du corpus une différence proposée par Fuentes Rodríguez (2009), selon laquelle *evidentemente* relève tant d'un registre soutenu et formel que d'un registre colloquial, ce qui ne serait pas le cas de *obviamente*, lequel relèverait exclusivement d'un registre soutenu.

Références bibliographiques

Real Academia Española (2013) : *Corpus del Diccionario histórico de la lengua española* (CDH) [en ligne]. <https://apps.rae.es/CNDHE> [dernière consultation : 01/02/2023]

Real Academia Española : Banco de datos (CREA) [en ligne]. *Corpus de referencia del español actual*. <https://www.rae.es> [dernière consultation : 01/02/2023]

Fuentes Rodríguez, C. (2009). *Diccionario de conectores y operadores del español*. Madrid : Arco/Libros.

Hummel, M. (2012). *Polifuncionalidad, polisemia y estrategia retórica. Los signos discursivos con base atributiva entre oralidad y escritura*. Berlin, Boston : De Gruyter.

Hummel, M. (2018). Romance sentence adverbs in *-mente*: Epistemic mitigation in synchrony and diachrony. *Linguistik Online*, 92.5.

Real Academia Española : *Diccionario de la lengua española*, 23.ª éd., [version 23.6 en ligne]. <https://dle.rae.es> [dernière consultation : 01/02/2023]

Sánchez Jiménez, S. U. (2013). La evolución de algunos adverbios evidenciales: *evidentemente*, *incuestionablemente*, *indudablemente*, *naturalmente*, *obviamente*. *In* M. P. Garcés Gómez (coord.), *Los adverbios con función discursiva: procesos de formación y evolución*. Madrid, Francfort : Iberoamericana, Vervuert, 239-73.

Torner Castells, S. (2016). Los adverbios evidenciales en español. *In* R. González Ruiz et al. (éd.). *La evidencialidad en español: teoría y descripción*. Madrid, Francfort : Iberoamericana, Vervuert, 261-76.

Corpus électroniques et l'enseignement de la traduction assistée par ordinateur

Ola Elghamry
Laboratoire Lidilem, Université Grenoble Alpes
Ola.El-ghamry@univ-grenoble-alpes.fr

Les nouvelles technologies ont eu un grand impact sur les pratiques de la traduction. Il existe une multitude d'outils qui se multiplient et se diversifient chaque année. L'ensemble de ces outils entre dans le champ de ce qui peut être appelé la traduction assistée par ordinateur (TAO), prise dans son acception la plus large. Dans ce cas, la TAO est envisagée comme « tout outil informatique mis à la disposition des professionnels de la traduction et facilitant leur travail. » (Frérot et Karagouch, 2016)

On a atteint aujourd'hui un résultat de traduction réalisée par la machine d'une haute qualité quasi-humaine. Pourtant, certains enseignants bannissent l'utilisation de la traduction automatique en cours de traduction, considérant que c'est une sorte de tricherie. Il s'agit là d'un déni d'une réalité qui évolue si l'on prend en considération le marché de la post-édition de traduction automatique. Selon le rapport de ELIS 2023 (*European Language Industry Survey* https://elis-survey.org/wp-content/uploads/2023/03/ELIS-2023-report.pdf) « La traduction automatique domine la liste des tendances » dans le marché de la traduction.

Si on peut se passer d'utiliser certaines technologies dans le domaine de l'acquisition de la langue, ceci se fera difficilement dans le marché de la traduction qui fait preuve d'une grande compétitivité et qui exige une haute productivité. Beaucoup de questions s'imposent donc dans le champ de la traductologie: Comment former les traducteurs d'aujourd'hui avec la technologie de demain? Comment éviter le hiatus entre la formation académique et professionnelle? Comment créer des ponts qui mèneront les apprenants vers le marché actuel de la traduction? Est-ce qu'on pourra un jour se passer d'un traducteur humain? Certainement oui s'il ne s'adapte pas au monde en constante évolution.

En connaissant mieux la technologie derrière la traduction automatique, le futur traducteur comprendra que la machine est essentiellement enrichie par le travail du traducteur : « Google Traduction et Deepl réussissent d'ailleurs à développer leurs systèmes de traduction neuronaux après que des utilisateur•rice•s aient alimenté et bonifié leurs bases de données pendant des années » (Simard, 2021)

Linguistes et informaticiens unissent leurs efforts pour exploiter ces « bases de données », ces mines d'or que constituent ces corpus électroniques qui existent en abondance et ne cessent de susciter de nouveaux outils pour les explorer et les analyser. Mona Baker est une pionnière dans le domaine de l'utilisation des corpus pour des recherches en traduction ouvrant la voie au «Corpus-Based translation studies » (Baker, 1998) ou en Français *La Traductologie de Corpus* (Loock, 2016). Les deux termes décrivent l'alliance entre la linguistique de corpus et la traductologie.

Dans le cadre de cette communication nous nous intéressons au rôle crucial que jouent les corpus électroniques au service de la traduction. Il s'agit notamment de corpus parallèles bilingues ou multilingues qui sont à la base de presque tous les moteurs de traduction ainsi que les concordanciers en ligne qui se sont multipliés récemment pour devenir une ressource importante pour le traducteur (par exemple Linguee, Reverso, Tradooit). Un corpus parallèle est : « un ensemble de textes accompagnés de leurs traductions dans une autre langue. » (Kraif, 2014). Nous appréhendons les corpus parallèles en tant qu'« outil pédagogique » (Frérot et Pecman, 2021). En formant aujourd'hui des traducteurs, il devient indispensable d'une part de former les étudiants à l'utilisation de ces outils et d'autre part de profiter de ces corpus dans l'apprentissage même des techniques de la traduction. Vue que « les corpus peuvent apporter des réponses que d'autres outils n'apportent pas, et peuvent également permettre des gains

importants en ce qui concerne la qualité des textes traduits, s'agissant notamment de l'amélioration de la fluidité et du caractère naturel de la langue » (Loock, 2016 :103). Le cadre européen des compétences de l'EMT de 2022 (European Master in Translation) est un référent de compétences en traduction en dépit de la langue utilisée, vient appuyer l'importance de ce type de formation dans la compétence numéro 16 : « Utiliser efficacement les moteurs de recherche, les outils basés sur les corpus, les outils d'analyse de texte et les outils de TAO. »

Nous présenterons ici un retour d'expérience pédagogique qui s'est déroulée dans le cadre d'un cours intitulé « Outils de traduction » pour 30 étudiants en 1ère année de licence au Département de Langue et de Littérature Françaises à la faculté des Lettres, à l'Université d'Alexandrie. L'objectif du cours vise à former les apprenants aux technologies de la traduction professionnelle. Le cours s'est étalé sur 12 semaines, 4h par semaine, 2h de théorie et 2h de pratique. Le contenu du cours s'est appuyé sur l'utilisation des technologies existantes allant de la simple utilisation des dictionnaires électroniques pour résoudre des problèmes linguistiques allant jusqu'à la constitution de mémoire de traduction définie entant qu' « une banque de données dans laquelle sont enregistrées, sous forme d'unités de traduction, un texte source et le texte cible correspondant. Cette mémoire qui constitue la base des logiciels de TAO permet de retrouver de façon automatique les passages déjà traduits. » (Torillas,2009). Le cours a également abordé la traduction automatique notamment le MTPE (*Machine Translation Post-Edition*) ou PE en français (la post-édition) qui consiste à corriger le résultat de la machine par un traducteur humain.

Nous focaliserons notre retour d'expérience sur deux activités d'apprentissage : la constitution de mémoire de traduction (corpus parallèle) et l'évaluation de la traduction automatique (corpus automatique). La première activité a été un apprentissage par projet et par groupe ayant pour but la formation de l'apprenant en traduction à constituer une mémoire de traduction à partir de corpus parallèles en utilisant un aligneur. Web Align[44] un logiciel d'alignement, développé par Olivier Kraif à l'université de Grenoble Aples, a été utilisé pour aligner les corpus et créer des mémoires de traduction. Cette activité vise deux objectifs : la formation à la technique, aux notions mises en pratique et la constitution d'un lexique qui améliorera leurs compétences langagières pour la paire de langues FR-AR. Il s'agit de maîtriser le savoir-faire de l'utilisation des corpus à des fins professionnelles.

Le libellé de l'activité est de constituer un glossaire et une mémoire de traduction à partir de corpus parallèle (FR-AR) en domaine spécialisé autour de 4 thématiques suggérées. La classe est divisée en groupe de 3-5 étudiants. Le résultat attendu est une présentation du travail effectué selon un modèle de rapport avec des rubriques fixes à remplir, comportant une réflexion sur leur travail et des livrables constitués d'une mémoire de traduction et d'un lexique.

Parmi les compétences du traducteur selon le référentiel EMT 2022, dans la rubrique Compétence stratégique, méthodologique et thématique, la compétence no 14 précise «Postéditer un produit de MT à l'aide de guides de style et de glossaires terminologiques pour veiller au respect des normes de qualité dans les projets de traduction automatique augmentée » La traduction automatique a été abordé au cours en présentant un aperçu historique des techniques sous-tendant la TA. Pour l'activité, le travail a consisté en une analyse critique du corpus de traduction automatique généré par un métamoteur de traduction *Memsource*[45], désormais *Phrase.*

L'objectif de l'activité vise à développer leurs compétences de critique de traduction pour s'entraîner à l'activité de la PE et à comprendre les capacités ainsi que les limites de la TA. La critique s'est faite en suivant des critères bien définis. Ceux-ci ont été établis en ayant recours à des réviseurs professionnels de traduction qui utilisent des critères fixés par les agences de traduction pour la révision de traducteurs humains. Nous avons conçu une grille d'analyse à partir de ces critères et il fallait les appliquer aux traductions générées par les moteurs de traduction. Nous avons seulement travaillé la version (FR-AR)

[44] http://phraseotext.univ-grenoble-alpes.fr/webAlignToolkit/

[45] https://cloud.memsource.com/web/login/auth

pour deux raisons : l'Arabe étant la langue maternelle des apprenants, c'est la langue cible qu'ils utiliseront en traduction professionnelle et pour détecter les erreurs plus facilement; ce qui n'est pas évident pour des étudiants en 1ère année même pour la langue maternelle. En faisant cette analyse, les apprenants ont bien saisi les capacités ainsi que les limites de la TA. En d'autres termes, ils auront intérêt à développer en premier lieu leurs compétences en traduction sans lesquelles ils ne pourront pas détecter les failles d'une TA et les corriger, et c'est dans ce sens que leur intervention va prédominer dans l'avenir de la traduction professionnelle.

En se basant sur les rapports rendus pour la constitution de corpus et l'évaluation de l'exercice de post-édition, nous présenterons les résultats de ces deux activités et détaillerons les difficultés rencontrées, ainsi que les savoir-faire et compétences professionnelles acquises. Nous verrons comment les étudiants parviendront à situer leur rôle par rapport à des outils et des ressources électroniques qui ne cessent de révolutionner les pratiques.

Références bibliographiques

Anthony Pym (2013). Translation Skill-Sets in a Machine-Translation Age, Meta, 58(3), pp. 487–503.

Baker, M. (1998). Réexplorer la langue de la traduction : une approche par corpus. *Meta*, *43*(4), 480–485.

Fouad, Maali (2014). La traduction spécialisée par corpus bilingues alignés : Le cas de la traduction arabe, Actes du colloque international LA-LEA, *Langues Culture et professionnalisation dans un contexte mondialisé,* Poitiers.

Frérot Cécile et Lionel Karagouch (2016). Outils d'aide à la traduction et formation de traducteurs : vers une adéquation des contenus pédagogiques avec la réalité technologique des traducteurs , *ILCEA*.

Frérot, C., Pecman, M. (2021). Introduction à la place des corpus et des outils numériques pour l'étude des langues de spécialité In : *Des corpus numériques à l'analyse linguistique en langues de spécialité* [en ligne]. Grenoble : UGA Éditions.

Kraif, Olivier (2014). *Corpus parallèles, corpus comparables : quels contrastes ?*. Synthèse d'HDR. Université de Poitiers.

Kübler Natalie (2011). Working with Corpora for Translation Teaching in a French-speaking setting , A. Frankenberg-Garcia, L. Flowerdew & G. Aston (dir.), *New Trends in Corpora and Language Learning*, Londres : Bloomsbury, 62-80.

Kübler,Nathalie, Alexandra Mestivier, Mojca Pecman (2018). Teaching specialised translation through corpus linguistics: quality assessment and methodology evaluation by experimental approach. *META : Journal des traducteurs / Meta: Translators' Journal* , Les presses de l'Université de Montréal, 63 (3), pp.806-824.

Lacour, P., Bénel, A., Eyraud, F., Freitas, A. & Zambon, D. (2010). TIC, collaboration et traduction : vers de nouveaux laboratoires numériques de translocalisation culturelle. *Meta*, *55*(4), 674–692.

Loock, Rudy (2016). *La traductologie de corpus*, Septentrion.

Torrellas Castillo, M. (2009). Corpus bilingues massifs et mémoires de traduction : la version espagnole des textes juridiques de l'UE. *Revue française de linguistique appliquée*, XIV, 83-92.

Valérie Simard (2021). La traduction à l'ère du digital labor in www.revue-ouvrage.org.

Zanettin Federico (1998). Bilingual Comparable Corpora and the Training of Translators, *Meta*, vol. 43, n° 4, p. 613-630.

The RTBF Corpus : a dataset of 750,000 Belgian French news articles published between 2008 and 2021

Louis Escouflaire[1,2], Jérémie Bogaert [3]
Antonin Descampe [1] et Cédrick Fairon [2]
[1] CENTAL UCLouvain , [2] ORM UCLouvain , [3] ICTEAM UCLouvain ,
(louis.escouflaire/jeremie.bogaert/antonin.descampe/cedrick.fairon)@uclouvain.be

Introduction

News corpora provide a large and diverse representation of written language that is essential for various fields of linguistics, such as lexicology, discourse analysis, sociolinguistics, and for training language models in NLP (Tognini-Bonelli, 2021). Their diachronic aspect allows for an investigation of language evolution over time and of its back and forth influence on society, making them valuable resources for linguistic research and teaching (Hilpert & Gries, 2016). Likewise, in media and journalism studies, news corpora are important both for qualitative and quantitative research (Conboy, 2007).

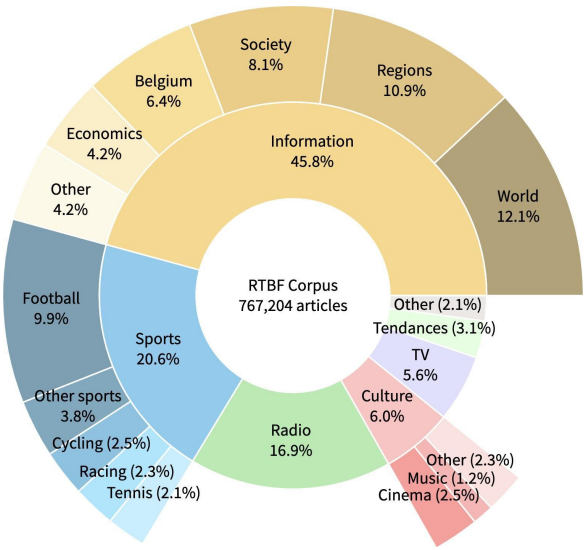


Figure 1 – Article distribution per feed (in percentages).

In this paper, we introduce the RTBF Corpus, a large diachronic corpus of 767,204 Belgian French news articles published between 2008 and 2021 by the Belgian public service media RTBF. We present the contents and structure of the corpus, along with the different layers of metadata available for each text. We also describe the three different versions of the articles available in the corpus (depending on the cleaning and preprocessing steps applied to the text). The RTBF corpus is freely available online in CSV format [46], for research and teaching purposes only.

The RTBF Corpus

There exist few open-source French news corpora of considerable size and interest. The *Est Républicain* Corpus, released in 2009, is to our knowledge the largest freely available French news corpus

[46] https ://dataverse.uclouvain.be/dataset.xhtml ?persistentId=doi :10.14428/DVN/PEVSSI

with around 149 million words (Gaiffe & Nehbi, 2009). The articles it contains were published between 1999 and 2003 and between 2006 and 2011, and cover mostly local news of the Eastern part of France (Seddah et al., 2012), which may restrict the variety of topics mentioned in the corpus.

The RTBF Corpus is a freely available Belgian French news corpus, like the *Est Républicain* corpus. However, the RTBF Corpus contains more than 214 million words, which makes it 1.4 times larger. RTBF is a national media which covers all kinds of topics, among which international and Belgian news, sports and culture. The corpus is chronologically continous (unlike the *Est Republicain* corpus) : the articles were published between 2008 and 2021, allowing for longitudinal investigations over the whole 2010 decade and more, and for analyses on news coverage of recent events (Escouflaire et al., 2022 ; Escouflaire et al., 2023). The corpus can also be a useful resource for experiments involving machine learning or requiring large amounts of data, such as news genre classification. Such large-scale resources are limited for the French language, making this corpus a useful asset.

The news director of RTBF officially handed us the rights over this dynamic corpus, which will be updated regularly with new articles published on the RTBF website. The corpus will be available online for free downloading, if the user agrees to follow the terms of use which are attached to the data. In short, the corpus may exclusively be used for research or teaching purposes (no commercial usage is permitted), it must be exploited in an ethical manner, and its user must inform RTBF of any planned publications related to the results and conclusions obtained from the corpus, and the media has the ability to request relevant comments and observations made by them to be included.

Corpus description

RTBF (*Radio-télévision belge de la Communauté Française*) is the public service broadcasting organization for the French-speaking community of Belgium. As a public service media, it is directly funded by the Belgian government and has three main missions : inform, educate, and entertain a public as large as possible in the French-speaking Belgian community. Besides operating television channels and radio stations, RTBF also operates a news website since 2008, on which web-only press articles are published daily. Through scientific collaboration with RTBF, we received access to the entirety of the web articles published on their website from 2008 to 2021. As shown in Figure 2, the publication rate has changed over the years, likely due to editorial movements inside the media. As of January 2023, between 1,000 and 1,500 articles are uploaded every week on their website.

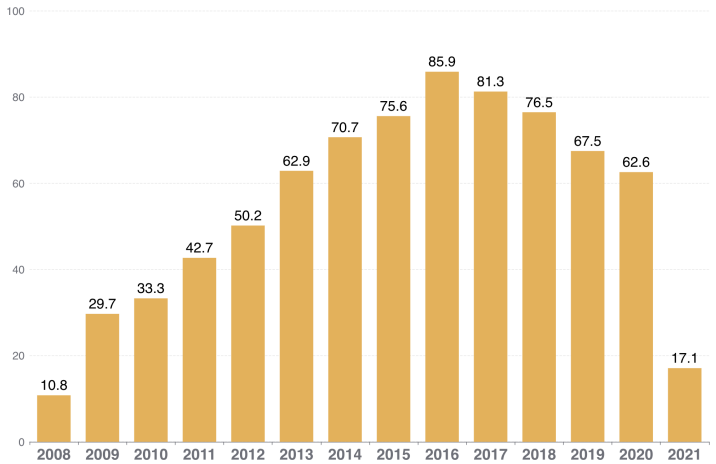


Figure 2 – Article distribution per year (in thousands of articles).

The corpus contains 767,204 articles, for a total of 214,072,620 words. The mean amount of words per article is 279. The full text of each article is in the *text* column of the corpus, along with seven columns consisting of different pieces of metadata attached to the article : *ID*, *title*, *publication date*, *signature*, *feed*, *category* and *keyword*. We present the contents of all columns in more detail :

ID : The article's internal ID, which is a number containing between 3 and 8 digits generated by the media source. This ID can be exploited to easily find its original page and layout on the RTBF website. A straight Google search with the format "*RTBF [ID]*" will return the article's link as the first result.

- ***title*** : The article's title, with an average length of 10 words. The lead paragraph and subtitles of the article are part of the main *text* column.

- ***publication date*** : The day on which the first version of the article was published on the RTBF website. The content of some articles may have been updated after their publication, but this information is not present in the corpus.

- ***signature*** : The name of the journalist who wrote the article, the radio or TV program it was originally derived from (cfr. *feed) further*, or the original source which the information is attributed to. About 43% of all articles in the corpus are signed solely or in collaboration with *Belga* or *AFP* (Agence France-Presse), two renowned press agencies delivering straight news. A great number of press agency dispatches are published daily on the RTBF website, with or without enrichments made by RTBF journalists. 1% of all articles were signed by a person or an entity who signed only one article in the corpus.

- ***feed*** : The structural and topical section of the RTBF website in which the article was published. The entire corpus is divided into 7 feeds : *Information*, *Sports*, *Radio*, *Culture*, *TV*, *Tendances* and *Other*. Their percentage distribution in represented in the inner circle of Figure 1. The *Information* feed, which is also the default feed on the RTBF website, contains mostly news belonging to the *information* genre of journalism (Grosse, 2001), i.e. content that is usually considered objective and factual by the source and by readers. Most press agency dispatches are found in this feed. However, the *Information* feed also contains 5,000 articles belonging to the *opinion* genre, i.e. op-eds or columns (those articles are marked with the *chronicles* or *opinions* category tag). The *Sports* and *Culture* feeds also include some amount of opinionated articles (*chronicles)*, representing respectively 416 (0.2%) and 2,800 (6%) of their articles. Then, the *Radio* and *TV* feeds contain articles which were derived from programs broadcasted on RTBF radio or television channels. *Tendances* contains lifestyle articles (e.g. fashion, food, technology, health, travel). The *Other* feed groups articles that were not classified into the 6 other feeds, among which many press releases and weather reports.

- ***category*** : Further classification related to the article's topics. Categories can help to group articles into more specific themes, as shown on the outer circle of Figure 1. Categories with small amounts of articles (e.g. *Basketball*, *Gaming*) were not represented on Figure 1, but were instead grouped into the *Other* category of the feed (*e.g. Other sports*). For the *Radio* and *TV* feeds, categories (not shown on Figure 1) are related to the specific TV channel or radio station from which the broadcast-related articles were derived (e.g. "Matin Première").

- ***keyword*** : Word or phrase attributed to some articles by the journalist or source who wrote them to further annotate their content (e.g. *Europe*, *environment*, *Internet*). About 28,5% articles in the corpus were assigned a keyword.

Corpus cleaning and preprocessing

Three different versions of the article's full text are available in the corpus. Each version contains the text at different stages of data cleaning and preprocessing :

- *HTML* text : The raw version of the text directly received from the RTBF, containing HTML tags. This version of the text can be used for research dealing with metatextual information (e.g. words highlighted in bold or italic, headings).

- *cleaned* text : The version of the text after multiple steps of data cleaning. To get this version, we removed HTML tags and fixed bugs related to text encoding of special characters. Some

metatextual elements are still included in this version, such as signatures added at the end of some texts and links referring to other articles of the website.

- *preprocessed* text : The version obtained after multiple preprocessing steps. Most signatures at the end of articles (e.g. *with Belga*) and links to other articles (e.g. *Read also : ...*) were systematically removed from the articles. Numbers and URLs were replaced by placeholder tokens, <NUM> and <URL>. This version should be used carefully, as some elements that may be relevant for some experiments could be missing from the text. Those preprocessing steps were initially derived and designed for the creation of the InfOpinion corpus (*Bogaert* et al., *2023*), a subcorpus of the RTBF corpus fit for a text classification task.

Notes

The RTBF Corpus was initially created in the context of a PhD program carried out at UCLou- vain, in partnership with RTBF. The research is funded by FRS-FNRS (Belgian National Fund for Scientific Research) and conducted by PhD student Louis Escouflaire at ILC (Institute for Language and Communication) under the guidance of Antonin Descampe (Observatory for Research on Media and Journalism) and Cédrick Fairon (Center for Natural Language Processing).

References

Bogaert, J., Escouflaire, L., de Marneffe, M.-C., Descampe, A., Fairon, C. & Standaert, F.-X. (2023). TIPECS : A corpus cleaning method using machine learning and qualitative analysis. *International Conference on Corpus Linguistics 2023 (JLC23).*

Conboy, M. (2007). *The language of the news.* London : Routledge.

Escouflaire, L., Descampe, A., Fairon, C. (2022). L'évolution de la subjectivité linguistique dans le journalisme web du XXIe siècle : analyse d'un corpus belge francophone d'articles de 2010 à 2021. *JADT 2022 : 16th International Conference on Statistical Analysis of Textual Data.* Naples, Italy.

Escouflaire, L., Descampe, A., Lits, G., Fairon, C. (2023). Analyzing the semantic evolution of bias in French news articles using word embeddings. *Digital Humanities Benelux 2023.* Brussels, Belgium.

Gaiffe, B. & Nehbi, K. (2009). Le corpus de l'Est Républicain. *Technical report, Atilf.* http ://www.cnrtl.fr/corpus/estrepublicain/.

Hilpert, M., & Gries, S. T. (2016). Quantitative approaches to diachronic corpus linguistics. The Cambridge handbook of English historical linguistics, 36-53.

Küppers, A., & Ho-Dac, L. M. (2011). Un corpus de presse francophone pour l'étude de l'impact d'Internet sur les pratiques langagières. *CJC Praxiling : Corpus, données, modèles : approches qualitatives et quantitatives.*

Seddah, D., Candito, M., Crabbé, B., & Anguiano, E. H. (2012). Ubiquitous usage of a broad coverage French corpus : Processing the Est Républicain corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3249-3254.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at work.* Amsterdam/Philadelphia : John Benjamins.

TIPECS : A corpus cleaning method using machine learning and qualitative analysis

Louis Escouflaire[1,2], Jérémie Bogaert [3]
Antonin Descampe [1] et Cédrick Fairon [2]
[1] CENTAL UCLouvain , [2] ORM UCLouvain , [3] ICTEAM UCLouvain ,
(louis.escouflaire/jeremie.bogaert/antonin.descampe/cedrick.fairon)@uclouvain.be

1. Introduction

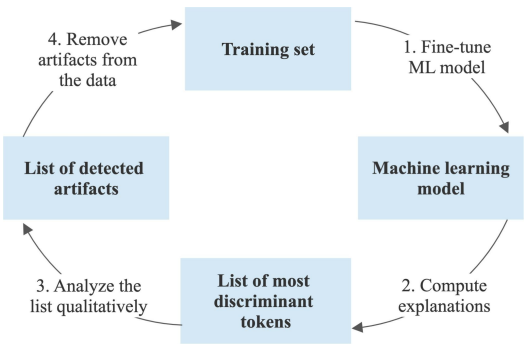
We present TIPECS (**T**rain, **I**nfer **P**redictions, **E**xplain, **C**lean and **S**tart again), a corpus cleaning method relying on a mixed approach between machine learning and manual analysis. The aim of our dataset cleaning approach is to remove tokens or segments that are considered as discriminant features by a classification model trained on a given dataset for a given task, but that cannot be generalized to other similar tasks or datasets. Such elements are referred to as *artifacts* (Gururangan et al. [2018]). For example, when classifying news articles as opinion or information, some press agencies signatures, such as Belga, will be clear indicators for the information class, but might not appear in other datasets. To make the dataset as valuable as possible for the task overall, these artifacts should be removed. We showcase TIPECS on a French information vs. opinion news classification task. The code for our method and the dataset created will both be available on GitHub[47].

2. The TIPECS method

TIPECS is a semi-automatic method to iteratively remove, from a given dataset, tokens that can be considered as artifacts for a given task. Figure 1 illustrates the process. The intuition behind the method is to use a powerful model overfitting on artifacts and biases to highlight them and remove them from the dataset. The method is applicable to all kinds of textual datasets and using various types of machine learning models. The different steps of the TIPECS method are the following :

- **T**rain : Feed the training data to the model to teach it to correctly classify the items.
- **I**nfer **P**redictions : Use the trained model to predict a class for some unseen items.
- **E**xplain : Use a token-level model explanation method to find the top-n tokens that influence the most the model's predictions (*n* can be determined by the user, default value is 20).
- **C**lean : If artifacts are found through qualitative analysis of the n-most discriminant tokens, add preprocessing steps to remove them from the dataset.
- **S**tart again : Repeat the process until the n-most discriminant tokens contains only items relevant to the task.

Some constraints must be respected to obtain valuable results with TIPECS. First, the machine learning model used has to be able to identify discriminant features for the classification task. Transformer models are therefore candidates of choice for our method, given their tendency to overfit on various biases and artifacts (Gururan- gan et al.



[47] Link to GitHub (anonymized).

[2018]). Second, the model has to be able to explain its predictions at token level, for example, with transformer models, using pertur- bation methods or attention probing (Bibal et al. [2022]). Finally, since the training, the prediction inference and the explanation steps have to becarried out at each cycle, these three steps should not be too computationally intensive.

FIGURE 1 – TIPECS framework to remove artifacts from a dataset.

3. Case study

3.1. The InfOpinion corpus

The InfOpinion corpus comes from the RTBF corpus, a collection of 750,000 news articles published by the Belgian French-speaking public service media RTBF (Radio-télévision belge de la Communauté Française) on their website between 2008 and 2021. This corpus was built for training and evalua- ting a classification model distinguishing between pieces from the journalistic *opinion* genre (such as editorials, commentaries, reviews), which are considered to be subjective, and pieces belonging to the *information* genre (press agency dispatches, news articles), meant to be more objective texts (Grosse [2001]). This binary categorization relies solely on the articles' annotation by the RTBF as either *opinion* or *information*.

The InfOpinion corpus consists of 10,000 articles published between 2012 and 2021 on the INFOR- MATION feed of the RTBF Corpus. It is a balanced dataset containing on the one hand 5,000 articles classified by the RTBF as *Op-eds* ("Chroniques") or *Opinions*, and on the other hand 5,000 articles randomly selected from the *Belgium*, *World*, and *Society* categories, which tackle similar topics to those discussed in the *Op-eds* category. Each article in the InfOpinion corpus has multiple layers of metadata, which are described further in (*Anonymized* et al., [*in press*]) : ID, title, publication date, signature, feed, category and keyword. An additional column is devoted to the article's genre (or class), which is either *information* or *opinion*.

3.2. Applying TIPECS

We showcase TIPECS by using it to preprocess the InfOpinion corpus with *information* vs. *opinion* classification in mind. We use CamemBERT, a French adaptation of the RoBERTa transformer model (Liu et al. [2019], Martin et al. [2020]) and an explanation method based on a variation of layer-wise propagation (LRP). The idea behind LRP is to explain a deep neural network prediction by assigning a relevance score to each input feature (i.e. token) by starting from the prediction and back-propagating to the input using conservation constraints (Bach et al. [2015]). We use Chefer et al. [2021]'s variation of the LRP method, which we modified to be able to work with CamemBERT's architecture. All other parameters remained unchanged from the original implementation. These settings allow us to obtain good and fast results with a single GPU (about 40 minutes per TIPECS cycle).

The InfOpinion corpus is first divided into a training set (80%), a development set (10%) and a test set (10%), all balanced between the *information* and *opinion* classes. We **train** the base version of CamemBERT on the training set during 2 epochs to classify *opinion* and *information* articles. We then **infer predictions** with the trained CamemBERT model on the development set and **explain** those predictions at token level with LRP by computing attention maps for the 1,000 articles in the development set. To avoid analyzing too specific tokens, we only take into account tokens with at least 10 occurrences in the dataset. For each token, words and punctuation signs included, we compute its mean relevance value across all the articles in which it appears. All tokens are then ranked by average relevance value. Table 1 shows the 20 most discriminant tokens obtained after the first iteration for both the *opinion* and *information* classes.

The next step consists in manually examining the two lists of the most discriminant tokens for the *opinion* and *information* classes, to find which elements are not relevant to the classification task. Then, we add preprocessing rules to **clean** the dataset by removing the unwanted tokens. Finally, after each TIPECS cycle, we **start again** : the CamemBERT model is fine-tuned on the newly cleaned version of the InfOpinion dataset, and all the steps in the TIPECS iterative process are repeated until no artifacts remain in the two lists of most discriminant tokens, i.e. until all tokens are judged potentially relevant and generalizable for explaining predictions in our task. In this case study, we stopped after 5 iterations.

In the first iteration, we identified multiple potential artifacts in the dataset related to the texts' structures, such as press agency or author signatures (e.g. *Belga*), links and references to similar articles (*Read also...*), and structural tokens (*#*). We therefore added a preprocessing step to keep only the text and remove the metadata surrounding it. We also replaced URLs with a unique tag (<URL>) to avoid model overfitting on the various hyperlinks present in the data. Similarly, we replaced all numbers and digits with a tag (<NUM>), to prevent the model from using meaningless numerical values during its predictions while still allowing it to highlight their presence or frequency.

Opinion				**Information**			
cinéaste *filmmaker*	provocation	film	foi *faith*	AFP</s>	"</s>	visant *aiming*	samedi *Saturday*
sinon *otherwise*	bref *in short*	actualité *news*	pari *bet*	Belga</s>	jeudi *Thursday*	indiqué *indicated*	Belga
latin	[	ah	imaginaire *imaginary*	précise *specifies*	expliqué *explained*	mardi *Tuesday*	bière *beer*
média *media*	puisqu'il *because he*	prenons *[we] take*	raconter *to tell*	pompiers *firefighters*	précisant *specifying*	mercredi *Wednesday*	concrètement *concretely*
#	cinéma *cinema*	incapable *unable*	décidément *decidedly*	rapporte *reports*	.</s>	précisé *specified*	ajouté *added*

Table 1 – Top-20 most discriminant tokens (based on LRP explanations) for the *opinion* and *infor- mation* classes after the first TIPECS iteration (from top to bottom, left to right).

Throughout the iterations of the process, we also observed that some topics were significantly more frequent in *opinion* articles than in *information* articles, in particular tokens associated with cultural topics (e.g. *cinema, film*). To address this, we modified the information dataset by replacing 10% of the articles with articles from the CULTURE feed of the RTBF corpus.

We found that text length was also playing a major role in the model's predictions. *Information* articles are on average shorter than *opinion* pieces. Since the inherent input limit for the Camem- BERT model is of 512 tokens, the end-of-sequence character (</s>) appeared more often after a final punctuation sign (., ! , ? or ...) in *information* articles, as seen in Table 1. On the contrary, it appeared more often in the middle of sentences for *opinion* articles (because the input texts were too long). To avoid this bias, we implemented a dynamic truncation preprocessing step allowing to cut long articles after the last complete sentence (i.e. ending with ., ?, ! or ...) ending before the token limit. This way, the model can no longer discriminate articles based on the end-of-sequence character.

Table 2 shows the final lists of most discriminant tokens we obtained on the development set after 5 TIPECS iterations. Note that since both classes do not exactly treat the same topics, words such as *hospital* or *iPhone* appear in the lists. They are part of semantic fields inherently more prevalent in their respective class (*information* and *opinion*), and thus cannot be considered artifacts in this task, as they could generalize to other datasets. This case study shows how TIPECS can be used to identify artifacts in a classification task dataset to prune it from unwanted bias during subsequent experiments.

Opinion				Information			
verbe *verb*	latin	église *church*	imaginons *[we] imagine*	affiche *poster/displays*	indique *indicates*	samedi *Saturday*	métrage *footage*
puisqu'il *because he*	sondage *poll*	iPhone	humour *humor*	mardi *Tuesday*	jeudi *Thursday*	vendredi *Friday*	ajouté *added*
patrons *managers*	paraît *seems*	budgétaires *budgetary*	médiatique *media-related*	hôpital *hospital*	mercredi *Wednesday*	passagers *passengers*	précisé *specified*
conservateurs *conservatives*	articles	mots *words*	découvre *discovers*	indiqué *indicated*	située *located*	lundi *Monday*	ONG *NGO*
syndical *union*	expression	intéresse *interests*	déficit *deficit*	exposition *exhibition*	aérienne *aerial*	AFP	précise *specifies*

TABLE 2 – Top-20 most discriminant tokens (based on LRP explanations) for the *opinion* and *information* classes after 5 iterations of the TIPECS method (from top to bottom, left to right).

References

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R. Bowman, S. R., Smith., N. A. (2018). An- notation artifacts in natural language inference data. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguis- tics : Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics. doi : 10.18653/v1/n18-2017.

Bibal, A. Cardon, R. Alfter, D., Wilkens, R., Wang, X., François, T., Watrin., P. (2022). Is Attention Explanation ? An Introduction to the Debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics. doi : 10.18653/v1/2022.acl-long.269.

Grosse, E.-U. (2001). Evolution et typologie des genres journalistiques. Essai d'une vue d'ensemble. *Semen. Revue de sémio-linguistique des textes et discours*, (13).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa : A robustly optimized BERT pretraining approach. *arXiv preprint arXiv :1907.11692*.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de la Clergerie, E., Seddah, D., Sagot, B. (202).. CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219. doi : 10.18653/v1/ 2020.acl-main.645. arXiv :1911.03894 [cs].

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7) : e0130140.

Chefer, H., Gur, S., Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791.

Apports de corpus multimodaux distincts pour la didactique du « français tout court ». Dépasser la dichotomie oral/écrit dans les textes d'élèves.

Auphélie Ferreira[1], Arnaud Moysan[2]
[1]Lattice (UMR 8094), Université Sorbonne Nouvelle, [2]CLESTHIA (EA 7345), Université Sorbonne Nouvelle
auphelie.ferreira@sorbonne-nouvelle.fr, arnaud.moysan@sorbonne-nouvelle.fr

Introduction

La présente communication s'est donnée pour objectif d'investiguer trois corpus différents pour mieux comprendre un phénomène peu éclairé en didactique du français écrit : le rapport à l'oralité chez les enseignants de français dans le cadre de productions d'écrits scolaires. Nous situons d'emblée notre propos dans un entrelacs de tensions, à commencer par la distinction entre l'écrit et l'oral dans le milieu scolaire, révélatrice du rapport qu'entretient l'institution scolaire à ces deux aspects de la langue française. De cette tension résulte un rapport de domination dans lequel l'écrit et ses formes dominent nettement toute forme de l'oral (Barré-de-Miniac, 2000), ce que révèle d'ailleurs la lecture des Instructions Officielles. Gagnon et Benzitoun (2020) à la suite des travaux de Blanche-Benveniste (2013) ou Béguelin (2000), rappellent justement que la représentation de français parlé comme forme fautive, appauvrie du français a été écarté de l'enseignement.

Du point de vue de la réception de ces textes d'élèves corrigés, cette tension entre oral et écrit apparaît sans équivoque dans de nombreuses formules prescriptives qui sanctionnent l'utilisation d'une langue dite « orale », en opposition à une langue « écrite » commandée par le professeur et ce même lorsque la consigne d'écriture, qui amorce le texte à produire (Garcia-Debanc, 1996), appelle une forme relevant de l'oral comme ce peut être le cas pour un dialogue écrit. C'est sciemment que nous mettons ces termes entre guillemets en raison du fait qu'ils posent problème, et ce à plusieurs égards, notamment à la lumière des travaux de Söll (1974) et, plus récemment, de Koch et Oesterreicher (2001). Ainsi, nous envisageons le rapport entre l'oral et l'écrit comme un « continuum conceptionnel » distinguant « domaine de l'immédiat » et « domaine de la distance ». L'observation des copies met en exergue que les choix des formes linguistiques par les élèves semblent être motivées par de multiples facteurs (origine sociale des locuteurs, medium communicationnel, etc.) qui dépassent les attentes des enseignants dans le cadre des productions d'écrits.

Plusieurs questions animent notre démarche : comment se manifestent les marques de l'oralité relevant du « code graphique » mais comportant une « allure parlée » (Mahrer, 2019) dans un texte d'élève ? Quelles sont celles qui sont relevées par les enseignants et comment sont-elles traitées ? Quelles représentations sous-tendent ces pratiques professionnelles de la correction ?

Pour répondre à ces questions, nous interrogeons plusieurs corpus, à la fois écrits, oraux et visuels. Notre démarche se veut originale puisqu'elle met en lien des corpus qui ont été

élaborés pour répondre à des objectifs initiaux distincts. Cette étude montrera ainsi comment exploiter des corpus au-delà de leurs objectifs initiaux et quels apports et limites une telle démarche entraîne nécessairement.

Corpus et méthodologie

Corpus

Le corpus d'étude est constitué de deux types de corpus distincts : (i) un corpus de ressources scolaires composé d'écrits scolaires et de vidéos de séances d'écriture en classe (ii) un corpus oral rassemblant le MPF[48] (*Multicultural Paris French*) et la partie orale du CEFC[49] (Corpus d'Étude pour le Français Contemporain).

Le corpus sur lequel repose notre étude provient donc de plusieurs sources qui en font un corpus multimodal. Celui-ci inclut tout d'abord des copies d'élèves recueillies en contexte écologique durant l'année scolaire 2018-2019 auprès d'enseignants de français exerçant à Paris et en région parisienne. Ces copies ont fait l'objet d'une transcription et d'une annotation suivant le protocole du projet ANR- E:Calm[50] permettant l'exploration et la fouille automatique de corpus initialement manuscrits dont une partie est consultable dans Ecriscol[51]. Les textes produits par les élèves ont été réalisés durant des séances dédiées à l'écriture, en classe, sous le pilotage de l'enseignant référent, et dont les séances ont pu être filmées.

Ce corpus de ressources scolaires est complété de données orales issues du MPF et du CEFC. Le MPF rassemble des données orales recueillies dans l'Île-de-France multiculturelle (au total 1,2 millions de token). L'objectif initial des linguistes à l'initiative de ce corpus était de permettre une analyse du Vernaculaire Urbain contemporain (Gadet, 2017 : 46). Quant au CEFC, il a été constitué pour répondre au besoin d'un corpus de grande envergure pour l'étude du français tout court (Debaisieux et Benzitoun, 2021). Sa partie orale a été exploitée dans le cadre de cette recherche (soit environ 3,1 millions de token).

Méthodologie

Bien que ces deux corpus oraux semblent éloignés des écrits scolaires et des vidéos des séances d'écriture, ils ont rejoint notre corpus d'étude pour l'observation des phénomènes perçus comme oraux. Ils ont permis, d'une part, de vérifier si des phénomènes linguistiques tels que la négation à un élément *pas*, le détachement ou encore l'emploi de certains lexèmes sont effectivement présents, ou plus encore, spécifiques à l'oral. D'autre part, ils ont participé au dépassement de la dichotomie oral/écrit grâce à une catégorisation des enregistrements proposée dans le MPF selon des critères internes et externes (Moreno, 2017) sur un continuum entre distance et proximité reposant sur celui proposé par Koch et Oesterreicher (2001). Les catégories proposées par le MPF ont été transposées au CEFC. Ainsi, l'intégration de ces données orales servent une analyse en termes de « genres discursifs », ou encore selon le continuum de proximité ou immédiat communicationnel et à « sortir des mythes séparateurs » (Blanche-Benveniste, 2010).

[48] https://www.ortolang.fr/market/corpora/mpf

[49] https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/index.html

[50] http://e-calm.huma-num.fr/

[51] http://www.univ-paris3.fr/ecriscol-300509.kjsp

Cette première étude s'est concentrée sur les classes de troisième. Notre choix s'est porté sur cette année car d'une part, peu de recherches se sont intéressées à ce niveau et que d'autre part il renvoie à la fin de la scolarité du collège et du cycle 4 ce qui fait que les élèves ne sont certes pas des scripteurs experts, mais ils ne sont pas des débutants pour autant. L'exploitation des corpus s'est faite en différentes étapes. Une première étape s'est portée sur le corpus écrit. Elle a consisté à relever les interventions des enseignantes concernant le caractère oral dans l'écrit de l'élève. Au total, 150 copies ont été analysées de manière manuelle ou semi-automatique lorsqu'elles ont été transcrites et interrogeables avec le logiciel en ligne itrameur[52]. Dans une seconde étape, l'observation des phénomènes annotés dans les corpus oraux a permis de confirmer ou au contraire d'infirmer le caractère oral du phénomène annoté par l'enseignante et d'identifier s'il s'agissait de marques d'un patron de l'oral, éventuellement relevant du pôle de la proximité communicative. Enfin, 6 vidéos de 45 minutes chacune correspondant aux séances d'écriture sur lequel repose le recueil des copies d'élèves ont été exploitées Ces séances filmées complètent les commentaires formulés par les enseignants sur les copies. Les commentaires sur la consigne et son explicitation, les attentes concernant la production écrite et sur l'écriture en cours ont été relevés et nous ont permis d'observer ce qui, pour les enseignants, pouvait relever de la norme et de sa transgression de l'oral dans un écrit (scolaire).

Résultats

Les phénomènes linguistiques pointés par les enseignants portent principalement sur la segmenta- tion, la syntaxe et le lexique employés par les élèves. Bien que les remarques concernant la ponctuation soient nombreuses dans les textes, le niveau segmental sera mis de côté dans ce premier travail et ex- ploré dans une étude ultérieure. Nous focalisons ainsi notre étude sur des phénomènes syntaxiques tels que les *hanging topics*, les détachements, l'usage des adverbes de négation ou encore les usages lexicaux avec notamment l'emploi de termes qualifiés par les enseignants de « familier ».

Par exemple, nous avons constaté que l'emploi du *ça* est parfois accepté mais bien souvent corrigé par *cela* par les enseignants. *Ça* est jugé comme familier (*cf.* fig. 1) et inadapté quand bien même la consigne invite l'élève à produire un énoncé oral entre deux amis. L'emploi de *ça* à l'oral dans un tel contexte est pourtant attesté dans notre corpus oral. Par ailleurs, Guerin (2021) rappelle que *ça* ne doit pas être envisagé comme l'équivalent oral et/ou familier de *cela* ayant un sens procédural spécifique.

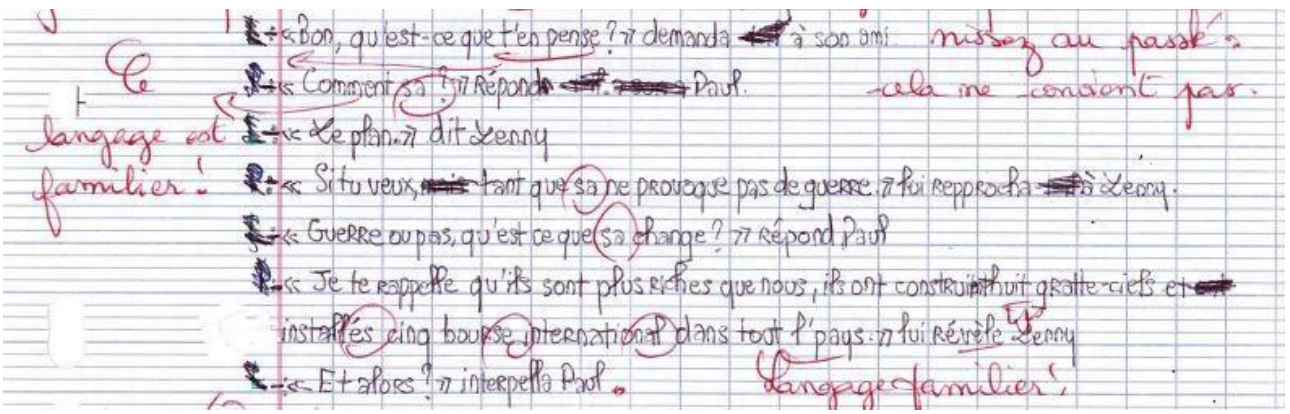


FIGURE 1 – Extrait d'une copie d'élève. Illustration de commentaires « langage familier » de la part d'un enseignant

En outre, certaines formes « orales » seraient plus acceptées que d'autres. Par opposition à *ça*, les *hanging topics* (Blasco-Dulbecco, 1999), définis comme des syntagmes adjoints à l'ensemble de la construction verbale sans reprise pronominale, ou les dislocations ne sont pas relevées par les en- seignants. Ces phénomènes décrits dans le cadre de productions orales, notamment par Raickovic et Skrovec (2020) travaillant sur le corpus ESLO, ont été relevés par ces mêmes auteurs dans les contextes de proximité communicative. Leurs emplois par les élèves, ce qu'illustrent les énoncés (1) et (2), donnent à voir leurs connaissances implicites[53] quant aux formes linguistiques employées à l'oral ou plutôt dans des contextes de proximité communicationnel.

(1) L'égalité entres les noirs et les blancs c'est quoi le problème ! (devoir1_E1)

(2) moi j'ai une famille, toi tu en a une je sais que tu veux la protégé (devoir2_E1)

Ainsi, en classe de troisième, certains semblent bien saisir ce qui est en jeu dans un échange et influence le recours à certaines formes linguistiques, toutefois ils se voient parfois corrigés. Il semble persister une confusion chez les enseignants entre « registre » de langue et ce qui relève d'un français parlé telle que le pointait du doigt Blanche-Benveniste (2013). Afin de mieux saisir la place de cette confusion aujourd'hui, une étude de plus grande ampleur devra être menée. Une pédagogie de l'oral pourra s'appuyer sur les corpus oraux et les travaux menés sur le français parlé, telle que le suggère Gagnon, de Pietro et Fischer (2017). Les corpus qui avaient des objectifs initiaux distincts ont été rapprochés afin de proposer cette étude qui, nous l'espérons, contribuera à rappeler la place importante des corpus dans la didactique du français, du français parlé et du français tout court.

Références bibliographiques

Barré-de Miniac, C. (2000). *Le rapport à l'écriture : aspects théoriques et didactiques*. Presses Universitaires du Septen- trion.

Blanche-Benveniste, C. (1988/2013). De quelques relations entre le lexique et la grammaire dans l'analyse du français parlé. In M.-N. Roubaud (dir.), *Langue et enseignement ; une sélection de 22 manuscrits de Claire Blanche- Benveniste (de 1976 à 2008)*, Tranel, 58, 165-171.

Blanche-Benveniste, C. (2010) *Le français : usages de la langue parlée*. Leuven, Belgique : Peeters.

[53] « Les connaissances implicites sont des connaissances dont l'individu n'a pas conscience, elles sont non verbalisables et donnent lieu à un sentiment puissant d'intuition, puisque l'apprenant n'est pas conscient de son savoir alors même qu'il fait preuve d'une capacité à l'utiliser. » (Nadeau et Fisher, 2011 : 3)

Blasco-Dulbecco, M. (1999). *Les dislocations en français contemporain, étude syntaxique*. Paris : Honoré Champion. Béguelin, M.-J. (dir.) (2000). *De la phrase aux énoncés : grammaire scolaire et descriptions linguistiques*. Bruxelles, Belgique : De Boeck-Duculot.

Debaisieux, J-M., Benzitoun, C. (2020). Orféo : un corpus et une plateforme pour l'étude du français contemporain. *Langages, 219*, 9-24.

Gadet, F. (2017). *Les parlers jeunes dans l'Île-de-France multiculturelle*. Paris : Ophrys. Garcia-Debanc, C. (1996). Consigne d'écriture et création. *Pratiques, 89*, 69-88.

Gagnon, R., Pietro, J.-F., Fisher, C. (2017). L'oral aujourd'hui : perspectives didactiques. In J.-F. de Pietro, C. Fisher, R. Gagnon, *L'oral aujourd'hui : perspectives didactiques*, 11-36, Namur, Belgique : Presses universitaires de Namur.

Gagnon, R., Benzitoun, C. (2020). Le français parlé comme objet d'enseignement ? Regards croisés d'une didacticienne et d'un linguiste. *Formations et pratiques d'enseignement en questions, 26*, 37-51.

Guérin, E. (2021). Une description fondée sur l'oral (?) penser ça sans cela. In P. Cappeau (dir.) *Une grammaire à l'aune de l'oral*. 146-154.

Koch, P., Osterreicher, W. (2001). Gesprochene Sprache und geschriebene Sprache / Langage parlé et langage écrit. G. Holtus, M. Metzeltin, Ch. Schmitt (éds), *Lexikon der Romanistischen Linguistik, Bd. I/2*, 584-627.

Mahrer, R. (2019). Parler, écrire : « continuum communicatif » et rupture matérielle. *Pratiques*, 183-184.

Moreno, A. (2017). Au-delà des genres de discours : le discours direct à travers les notions de proximité et de commu- nautés de pratique. *Cahiers de praxématique, 69*.

Nadeau, M., Fisher, C. (2011) Les connaissances implicites et explicites en grammaire : quelle importance pour l'ensei- gnement ? Quelles conséquences ? *Bellaterra Journal of Teaching & Learning Language & Literature 4*, 1-31.

Raickovic, L., Skrovec, M. (2020). « Syntaxe en interaction et variation diaphasique : l'exemple des dislocations dans ESLO2 ». *SHS Web of Conferences 78*, 14004.

Söll, L. (1974). *Gesprochenes und geschriebenes Französisch*. Berlin : Schmidt.

Quels indices langagiers pour mesurer les progrès d'élèves de maternelle ?

Oriane Gélin [1], Loïc Liégeois [1]
[1] Laboratoire CIREL-RECIFES, Université de Lille
[2] Université Paris Cité, LLF (UMR 7110) et CLILLAC-ARP (EA 3967)
oriane.gelin@univ-lille.fr, loic.liegeois@u-paris.fr

Introduction

Nous savons, notamment grâce aux travaux en sociologie, que les différences de compétences langagières pourraient provenir de contraintes socio-culturelles et, plus précisément, de dispositions intériorisées et variables en fonction du contexte familial (Bonnéry & Joigneaux, 2015 ; Lahire, 2019). Dans notre projet de doctorat, nous avons souhaité étudier la manière dont ces différences évoluaient, à l'école maternelle, entre la Petite Section et la Grande Section. Dans cet objectif, nous avons réalisé des enregistrements vidéo en classe afin de capter la restitution d'histoires par les élèves au cours d'une tâche de narration faisant suite à la lecture par l'enseignante d'un album de jeunesse.

Au cours de cette communication, nous souhaitons principalement présenter la réflexion que nous avons menée afin de déterminer les indices langagiers à retenir dans le but de mesurer l'évolution des compétences langagières dans les productions orales d'élèves de maternelle.

Dans un premier temps, nous nous arrêterons sur les problématiques liées au recueil, à la structuration et à l'analyse de notre corpus. Dans un deuxième temps, nous présenterons les indices d'évaluation retenus. Enfin, nous analyserons les premiers résultats issus de l'analyse de corpus. Nous montrerons notamment que les indices retenus permettent de mettre en évidence des trajectoires développementales globales tout en mettant en lumière les différences interindividuelles entre les sujets de l'étude.

Corpus et méthodologie

Ce travail repose sur les verbatims de deux cohortes d'élèves de maternelle, soit 23 élèves, dont 13 d'entre eux ont été suivis deux à trois ans entre la Petite Section (PS) et la Grande Section (GS). Ces élèves d'une école REP+ de Saint-Denis (93) sont issus de milieux sociaux contrastés bien que principalement de milieu populaire. Cela nous a semblé un lieu privilégié afin d'étudier les mécanismes impliqués dans les processus de construction ou de réduction des inégalités langagières. Ces élèves appartiennent à une unique classe de cycle (les trois sections de maternelle sont mélangées) dans laquelle les élèves demeurent avec la même enseignante pendant les trois années du cycle 1.

Le recueil du corpus s'est fait au moyen d'enregistrements vidéo de séances individuelles de type expérimental qui reposent sur la double stimulation (Vygotski, 1998 ; Brown et Ferrara, 1985). Le fait d'avoir enregistré les élèves individuellement leur offre un espace personnalisé dans lequel ils peuvent s'exprimer sans interruption et semble favorable à une mesure des compétences langagières en narration. Il s'agit en effet d'une tâche narrative proposée par

l'enseignante à chaque élève après la lecture d'un album en classe. Suite à cette lecture, la seule consigne proposée à chaque élève est de raconter ce qui vient d'être lu. Ce type de tâche, avec laquelle les élèves sont plus ou moins familiers en fonction des milieux sociaux, pourrait également nous permettre, au-delà des progressions langagières, de mesurer un éventuel « effet école » sur le développement des compétences orales. Au total, 153 enregistrements vidéo ont été réalisés entre 2016 et 2019. Les enfants enregistrés appartiennent à des niveaux différents (PS, MS et GS) mais font tous partie de la même classe. D'une année sur l'autre, les élèves franchissent un niveau tout en restant dans la même classe, ce qui a donc permis le suivi longitudinal de plusieurs d'entre eux. Ainsi, 8 élèves ont été enregistrés pendant leurs trois années scolaires d'école maternelle, 5 l'ont été sur deux années et 10 ont été enregistrés une année.

Nombre total d'élèves	Total de vidéos	Nombre de vidéos par sections			Répartition des élèves en fonction du nombre d'années de suivi		
		PS	MS	GS	1	2	3
23	153	46	52	55	10	5	8

Table 1 : Tableau récapitulatif du corpus

Une fois les vidéos recueillies, nous avons procédé à la transcription des interactions à l'aide du logiciel ELAN (Wittenburg et al., 2006). Si ce logiciel a été retenu pour la transcription, nous avons souhaité utiliser l'outil CLAN (MacWhinney, 2000) pour mener nos analyses quantitatives et qualitatives. En effet, CLAN se révèle particulièrement pertinent pour l'analyse des productions enfantines, notamment parce qu'il intègre un ensemble de fonctionnalités efficaces pour l'analyse de corpus en général (lemmatisation et étiquetage morphosyntaxique, calculs de fréquences et de concordances, par exemple) et pour l'étude du développement langagier en particulier (calcul de longueur moyenne d'énoncés ou de diversité lexicale, par exemple). Si la transcription a été effectuée avec ELAN, les conventions mises en œuvre correspondent à celles du projet CHILDES (MacWhinney, 2000). Une fois transcrites, les données ont été converties du format ELAN vers le format CHAT au moyen de l'outil Teicorpo (MoDyCo, 2016 ; Liégeois et al., 2015). Le corpus a enfin été étiqueté au niveau morphosyntaxique puis analysé à l'aide de la commande KIDEVAL (MacWhinney, 2022).

Notre regard s'est enfin porté sur le choix des indices langagiers à calculer pour répondre à notre objectif de mesure des progrès des élèves. Nous avons retenu les cinq indices suivants : la mesure de la diversité lexical (VOCD, Malvern et al., 2004), la mesure de la longueur moyenne des énoncés (MLU, Brown, 1973), ainsi que la fréquence des adverbes, conjonctions et pronoms relatifs. Les deux premiers indices ont été retenus car ce sont des indicateurs éprouvés de l'évolution langagière qui permettent aisément de comparer des productions d'enfants (Parisse et Le Normand, 2006). Si nous sommes conscients des limites de la mesure MLU, il nous a paru pertinent de le conserver dans notre contexte de production narrative, choix conforté par les résultats que nous vous présenterons ci-après.

Etant donné le contexte de notre étude, le choix d'analyser l'usage des adverbes, conjonctions et pronoms relatifs s'est fait au regard des programmes officiels de l'Education Nationale qui indique que les progrès des élèves « s'accompagnent d'un accroissement du vocabulaire et

d'une organisation de plus en plus complexe des phrases » (MEN, 2015a, p. 6). Plus précisément, le document d'accompagnement donne un tableau d'indicateurs de développement du langage parmi lesquels figurent les « Phrases complexes avec relatives, complétives, circonstancielle » (MEN, 2015b, p. 8). Nous avons ainsi décidé de retenir les catégories morphosyntaxiques introduisant les phrases complexes soit les adverbes, les conjonctions et les pronoms relatifs. Nous avons calculé la fréquence relative de chacune de ces formes en prenant en compte la fréquence totale des tokens de chaque transcription. Ainsi, nous éliminons les principaux biais possibles relatifs à la longueur de l'enregistrement, la couverture de chaque transcription (en tokens) et à la vitesse d'énonciation de chaque élève.

Résultats

À partir de nos cinq indices (VOCD, MLU, adverbes conjonctions et pronoms relatifs), nous avons tout d'abord procédé à une analyse en composante principale afin de savoir comment ces indices se comportaient individuellement et les uns vis-à-vis des autres, notamment afin de savoir si ces cinq indices pouvaient être résumés en un indice unique pour la suite de notre travail. La première dimension, axe qui explique à lui seule 35% de la variabilité, nous indique que tous les indices sont positifs et qu'ils s'orientent dans la même direction. Pour donner un exemple, cela signifie qu'un élève qui utilise en moyenne de nombreuses conjonctions, produit également des énoncés plus long et qu'il a une plus grande diversité lexicale. Les autres dimensions nous permettrons de prendre en compte des variations individuelles plus minoritaires vis-à-vis de ces cinq indices et d'expliquer des résultats qui pourraient s'écarter de ceux apportés par la première dimension.
Nous avons ensuite considéré cette analyse en composante principale en séparant les trois années (PS, MS, GS). Le graphique obtenu, dans lequel chaque point représente la moyenne d'un élève (PS en noir, MS ou rouge et GS en vert) illustre l'évolution de ces indices dans la direction de la dimension1 de notre analyse. Nous observons par ailleurs une harmonisation des scores entre élèves au fil des années.

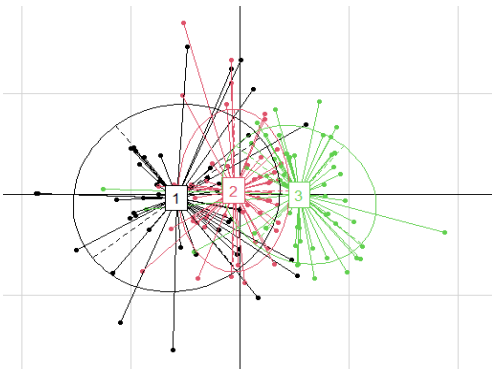


Figure 1 : PCA de la PS à la GS

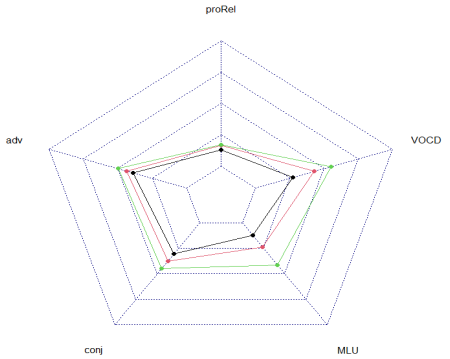


Figure 2 : Progression dans le temps des cinq indices

Quant aux indices en eux-mêmes, s'ils augmentent effectivement entre la PS et la GS, nous observons que la progression ne s'effectue pas dans les mêmes proportions en fonction des indices. Nous nous sommes enfin intéressés aux variations interindividuelles de progression en considérant dans un premier temps l'indice unique apporté par la dimension 1 de l'analyse en composantes principales puis, dans un second temps, chaque indice indépendamment. Si

tous les élèves voient leur diversité lexicale ainsi que la longueur de leurs énoncés augmenter, les résultats sont plus mitigés concernant les trois autres indices. Ce premier travail nous permet d'appréhender des cas particuliers d'élèves qui, contrairement à la tendance générale, voient leur usage des pronoms relatifs diminuer au fil du temps.

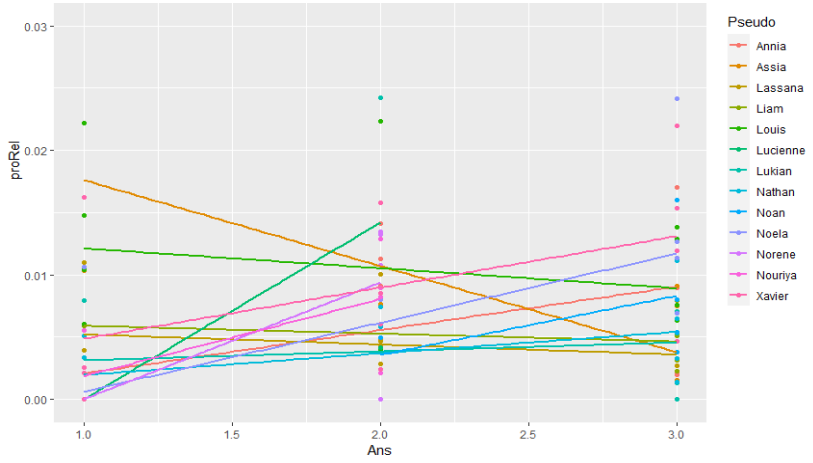


Figure 3 : Progressions individuelles de l'usage des pronoms relatifs de la PS à la GS

Si les résultats de ces premières analyses ne remettent pas en question la pertinence des indices choisis, ils invitent malgré tout à approfondir ces derniers, notamment par un retour au corpus pour des analyses qualitatives. Celles-ci permettront de déterminer les mécanismes en jeu dans la diminution, pour certains élèves, de l'usage de certaines formes. Les premiers résultats que nous venons de présenter dans cette proposition de communication seront donc approfondis et étayés par des analyses qualitatives au cours de notre communication.

Références bibliographiques

Bonnéry, S., & Joigneaux, C. (2015). Des littératies familiales inégalement rentables scolairement. *Le français aujourd'hui*, *3*, 23-34.

Brown, R. (1973). A First Language: The Early Stages* Harvard University Press. *Cambridge, Massachusetts*.

Brown, A. L., & Ferrara, R. A. (1985). Diagnosing zones of proximal development. *Culture, communication, and cognition: Vygotskian perspectives*, 273-305.

Liégeois, L., Etienne, C., Parisse, C., Benzitoun, C., Christian Chanard, C., (2015). Using the TEI as a pivot format for oral and multimodal language corpora. Text Encoding Initiative Conference and Member's meeting 2015, Oct 2015, Lyon, France.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Lawrence Erlbaum Associates.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development* (pp. 16-30). New York: Palgrave Macmillan.

Modèles, Dynamiques, Corpus - UMR 7114 (MoDyCo) (2016). teicorpo [Outil]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v1, https://hdl.handle.net/11403/teicorpo/v1.

Ministère de l'Education Nationale. (2015a). Programme d'enseignement de l'école maternelle, Bulletin Officiel spécial n°2 du 26 mars 2015

Ministère de l'Education Nationale. (2015b). Ressources maternelle – Mobiliser le langage dans toutes ses dimensions Parti I – L'oral – Tableaux d'indicateurs

Parisse, C., & Le Normand, M. T. (2006). Une méthode pour évaluer la production du langage spontané chez l'enfant de 2 à 4 ans. *Glossa (Paris)*, (97), 20-41.

Vygotsky, L. (1998). *The collected works of LS Vygotsky, Volume 5 : Child psychology (RW Rieber, Ed.; MJ Hall, Trans.)*.

Vygotsky, L. S., Van Der Veer, R. E., Valsiner, J. E., & Prout, T. T. (1994). *The Vygotsky reader*. Basil Blackwell.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 1556–1559.

La liaison dans un module d'ESLO-FLEU : mise en œuvre pour un cours de phonologie du français

Britta Gemmeke[1], Céline Dugua[2] et Flora Badin[2]
[1] Université de Siegen
[2] Laboratoire LLL-UMR7270, Université d'Orléans
britta.gemmeke@uni-siegen.de, celine.dugua@univ-orleans.fr, flora.badin@univ-orleans.fr

Introduction

Notre proposition s'inscrit dans le projet ESLO-FLEU (ESLO pour le FLE et la Linguistique dans l'Enseignement Universitaire), qui résulte de la collaboration entre le LLL à Orléans et le département des langues romanes à Siegen. A partir d'extraits du corpus de français parlé ESLO (*Enquêtes sociolinguistiques à Orléans*), le projet ESLO-FLEU a pour objectif de construire une ressource numérique sur la base de modules thématiques. L'enjeu est crucial ici : permettre aux créateurs de corpus oraux de rendre leurs données exploitables dans le cadre de cours de spécialité en linguistique, linguistique de corpus, sociolinguistique et didactique du FLE, et permettre aux enseignants de langues de s'approprier facilement ces corpus.

Dans ce travail, nous nous attachons à observer un phénomène phonologique particulier : les liaisons, qui consiste en la production d'une consonne de liaison entre deux mots (mot1 et mot2). Selon la nature des mot1 et mot2, les liaisons seront catégoriques (ou obligatoires), variables (ou facultatives), ou impossibles (ou interdites) (Delattre, 1947 ; De Jong, 1994). A l'instar du travail de De Jong (1994) sur le corpus d'Orléans, nous considérons dans cette étude seuls quatre contextes de liaisons catégoriques : entre déterminant et nom ou adjectif, entre pronom et verbe, entre verbe et pronom et dans quelques expressions figées. Les autres sont considérés comme variables, sauf quelques cas de liaisons impossible (après un adjectif au singulier, après la conjonction "et" par exemple).

La complexité de ce phénomène pour les apprenants se situe à plusieurs niveaux. Pendant que les apprenants moins avancés ont tendance à omettre la liaison obligatoire ou à réaliser la consonne de liaison sans resyllabification, les apprenants plus avancés et surtout les étudiants de français à l'université produisent des taux élevés de liaison correcte dans les contextes obligatoires (Mastromonaco, 1999 ; Thomas, 2004 ; Barreca, 2015 ; Pustka, 2015 ; Pustka et al., 2022). L'interaction avec l'orthographe semble aussi jouer un rôle important puisque la plupart des apprenants est confrontée à l'écrit dès le début du processus d'apprentissage. Cela peut poser problème surtout dans les cas où la consonne graphique ne correspond pas à la consonne attendue à l'oral (/t/ et non /d/ après "grand"). Même les apprenants avancés rencontrent des difficultés à distinguer le caractère obligatoire, variable ou interdit des liaisons. À propos des liaisons variables, on ne peut pas véritablement parler d'erreurs, mais on constate des différences de fréquences parfois considérables dans le sens où les apprenants avancés peuvent faire plus de liaisons variables que les locuteurs natifs (Pustka, 2015). Dans ce contexte, il semble que ce soit la variation stylistique et l'adaptation aux situations de communications qui posent le plus de problème aux apprenants.

Corpus et méthodologie

Corpus

Parmi les quatre modules thématiques ESLO-FLEU, le module « D'une situation à l'autre » créé autour de la variation diaphasique a été choisi pour l'étude de la liaison puisque son usage varie notamment selon les situations. Les types d'interactions qui y sont représentées sont diverses et relèvent d'une large gamme de situations allant de la conversation familière (tels que des repas, des discussions à la sortie du cinéma) aux discours publiques et académiques (des discours et des conférences par exemple). Ce module est également représentatif de la diversité générationnelle (diastratique), dans la mesure où il contient des locuteurs de toutes les tranches d'âges (de 15/25 à + de 65 ans) (Tahar et al., 2022). Il comprend 14 extraits issus de six situations d'ESLO pour une durée de 29 minutes.

Méthodologie

Dans le cadre du projet ESLO-FLEU, il est primordial de garder le lien entre transcription et son durant toute la durée de l'analyse. Ainsi, nous profitons du fait que les données initialement au format du logiciel *Transcriber*, soient converties dans un format adapté au logiciel TXM (Badin et al., 2021) qui intègre l'annotation, l'écoute de la donnée sonore, et la possibilité de recherches sur corpus (métadonnées incluses).

À partir des précédents travaux sur l'annotation de la liaison (Dugua et al., 2022), nous avons élaboré une requête complexe dans le concordancier de TXM qui repère tous les contextes potentiels de liaison dans le sous-corpus. Cette requête se fonde sur la forme graphique de deux mots consécutifs : le mot1 se termine par une consonne, le mot2 commence par une voyelle. Dans la requête nous excluons certains mots1 (*donc*, *et*...) et mots2 (*ah*, *oui*, *ouais*...) fréquents à l'oral, avant ou après lesquels la liaison est impossible. Cette définition large extrait de nombreux cas qui ne sont pas des contextes de liaison et que nous devrons éliminer au moment de l'annotation. L'annotation se fera alors sur le mot1 et ajoutera une propriété à ce mot selon le schéma d'annotation ci-dessous :

- La réalisation de la liaison : cette annotation permet d'informer, à partir de la consonne de liaison, si la liaison a été réalisée ou pas. Plus précisément, nous indiquons la nature de la consonne de liaison si la liaison est réalisée (|Z| dans "trois enfants", |N| dans "en été"), le codage |ABS| en cas d'absence de liaison dans un contexte possible ("c'est à", "dégoutant aussi") et le codage |Ø| pour les contextes impossibles.
- La classification du contexte en |catégorique|, |variable|, |impossible| ou |Ø| pour les cas repérés qui ne sont pas des contextes de liaisons.
- L'enchainement : ce codage concerne les cas de liaisons réalisées. Nous indiquons si les liaisons sont enchainées |Avec| ou non-enchainées |Sans|.

L'annotation des liaisons (524 occurrences) s'est faite selon un mode opératoire en trois temps. L'une des auteures a annoté l'ensemble du corpus, une autre a repris ces annotations pour les vérifier/modifier, et elles ont ensuite discuté sur les cas problématiques.

Résultats

Ce corpus comprend 405 contextes de liaisons qui se répartissent en 139 contextes catégoriques, 37 contextes impossibles et 229 contextes variables. Les taux de réalisation moyens dans chacun de ces contextes sont cohérents avec ce que l'on connait de l'usage des liaisons (Dugua et Baude, 2017). Même si ici les effectifs sont relativement faibles, on repère

un effet de la situation sur les taux de liaisons variables. Par exemple, on note 100% de liaisons variables réalisées dans les discours et 3.5% dans les repas. Nous reviendrons de façon détaillée sur ces résultats lors de la communication orale.

Outre l'intérêt de ce corpus pour rendre compte de l'usage des liaisons selon la variation diaphasique (possible grâce au module "D'une situation à l'autre"), nous avons créé un modèle de visualisation des annotations en liaison du module pour un usage d'enseignement. L'idée étant de proposer une présentation des transcriptions du module au format HTML dans lesquelles le son sera disponible, les contextes de liaisons seront surlignés et les annotations indiquées si besoin.

Mise en œuvre didactique

Une première mise en œuvre du module a été effectuée dans un cours de phonologie française à l'Université de Siegen (Allemagne) durant le semestre d'été 2023. Dans ce cours de première année de licence avec des étudiants issus de cursus différents (essentiellement formations des enseignants et sciences du langage), deux séances ont été consacrées au sujet de la liaison dans lesquelles les données ESLO-FLEU ont été utilisées. Ces séances visent principalement à introduire la liaison comme phénomène important de la phonologie française et à illustrer le caractère variable de la liaison dans sa dimension diaphasique. L'enseignante s'est servie des données et des annotations ESLO-FLEU notamment pour illustrer et alléger l'apport théorique et pour créer des exercices.

Dans la première unité de l'expérimentation, après une première familiarisation théorique avec le phénomène de la liaison, les étudiants ont travaillé en détail sur un extrait du module : ESLO2_CONF_1243_diamela-eltit. Il s'agit d'un extrait d'un discours académique d'une maître de conférences sur la biographie littéraire de l'écrivaine chilienne Diamela Eltit.

Le travail sur l'extrait s'est fait en trois parties :

1. Repérage des contextes de liaison possibles à partir de la transcription
2. Écoute de l'audio et indication pour chaque contexte possible si la liaison est réalisée ou pas
3. Précision de la nature de mot1 et mot2 et catégorisation du contexte de liaison en obligatoire, facultative ou interdite selon la classification de Delattre (1947)

Cette première unité aborde, avec un discours académique, une situation de communication formelle dans laquelle la classification classique et prescriptive de la liaison selon Delattre (1947) peut être bien appliquée.

La deuxième unité s'est concentrée sur le caractère variable de la liaison. D'abord, les travaux récents de la linguistique de corpus sur la liaison et leurs résultats (p.ex. Meinschäfer et al., 2015 ; Durand et Lyche, 2008) ont été présentés par l'enseignante et discutés en classe. Ainsi, la classification adoptée dans notre annotation a été élaborée en commun en cours. Ensuite, le corpus ESLO et le projet ESLO-FLEU ont été présentés.

Après une réflexion de potentiels questions de recherche à propos de la liaison dans des corpus, le sujet central étudié en cours était la variation diaphasique de la liaison. Une méthode d'apprentissage coopératif, parfois appelée « classe en puzzle » a été choisie pour pouvoir traiter le plus d'extraits possible et pour garantir l'application individuelle. Le détail du déroulement de la phase de travail sera présenté lors de la communication orale. Cette

séance visait à améliorer non seulement les connaissances des étudiants sur l'usage des liaisons mais aussi sur la linguistique de corpus.

Le travail sur corpus et l'authenticité des données constituent des pratiques innovantes et motivantes pour les étudiants. Nous espérons que la prédidactisation facilitera le travail des enseignants. Sur la base de ce premier retour d'expérience, la présentation orale sera l'occasion de rendre compte des apports de ce dispositif mis en place dans ESLO-FLEU et des difficultés qui subsistent à la didactisation de données issues d'un corpus oral.

Références bibliographiques

Badin, F., Liégeois, L., Thiberge, G. et Parisse, C. (2021). Vers un outillage informatique optimisé pour corpus langagiers oraux en vue d'une exploitation textométrique : le cas des interrogatives partielles dans ESLO. *Corpus*, 22. https://doi.org/10.4000/corpus.5752

Barreca, G. (2015). *L'acquisition de la liaison chez des apprenants italophones : des atouts d'un corpus de natifs pour l'étude de la liaison en français langue étrangère (FLE)* [thèse de coctorat, Paris Ouest Nanterre La Défense ; Università cattolica de Sacro Cuore di Milano]. https://www.theses.fr/2015PA100186

De Jong, D. (1994). La sociophonologie de la liaison orléanaise. Dans Lyche, C. (éd), *French generative phonology: retrospective and perspectives*, Association for French Language Studies, 95-129.

Delattre, P. (1947). La liaison en français : tendances et classification, *The French Review*, 21.2, 148-157.

Dugua, C. et Baude, O. (2017). La liaison à Orléans, corpus et changement linguistique : une première étude exploratoire. *Journal of French Language Studies*, 27.1, 41-54.

Dugua, C., Badin, F., Fallon, B. et Baude, O. (2022). L'usage des liaisons lors de lectures partagées – Une étude exploratoire à partir du module "Livres pour enfants" d'ESLO, *SHS Web of Conferences* 138, 09006, CMLF.

Durand, J. et Lyche, C. (2008). French liaison in the light of corpus data, *Journal of French Language Studies*, 18.1, 33-66.

Heiden S., Magué J.-P. et Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*, Rome, Italie, 1021-1032. http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf

Mastromonaco, S.M. (1999). *Liaison in French as a second language* [thèse de doctorat, University of Toronto]. https://hdl.handle.net/1807/12768

Meinschäfer, J., Bonifer, S. et Frisch, C. (2015). Variable and invariable liaison in a corpus of spoken French, *Journal of French Language Studies*, 25.3, 367-396.

Pustka, E. (2015). Die Liaison im Fremdspracherwerb: Eine Pilotstudie zu Münchner Lehramtsstudenten, *Bulletin VALS-ASLA*, 102, 43–64.

Pustka, E., Heiszenberger, E. et Hartmann, F. (2022). Pronunciation in Progress: A longitudinal study of the development of obligatory liaison in French as a foreign language, *Radical: A Journal of Phonology*, 3, 45-88.

Tahar, C., Skrovec, M. et Badin, F. (2022). Notice d'indexation thématique du sous-corpus ESLO-FLEU.

Thomas, A. (2004). Phonetic norm versus usage in advanced French as a second language, *International Review of Applied Linguistics in Language Teaching*, 42.4, 365-382.

Progressive forms vs. “en train de” in En/Fr Human and Machine Translation

Daniel Henkel[1]
[1] TransCrit, Université Paris 8 Vincennes-St. Denis
daniel.henkel@univ-paris8.fr

1. Introduction

Bilingual corpora are traditionally described as “comparable” corpora, consisting of similar original texts in two languages, and “parallel” corpora made up of original texts in a source-language and texts translated into a second target-language (McEnery & Xiao 2007).

Since it was first defined by Johansson in 2007, the “bidirectional” corpus model, combining both comparable and parallel corpora, has taken on increasing importance (cf. Frankenberg-Garcia 2009, Zanettin 2011, *inter alia*). The bidirectional approach allows for comparisons on multiple levels:

• between comparable corpora of original texts, to establish benchmarks for each language without any interlinguistic interference;

• between subcorpora of target-texts and original texts in the same language, to evaluate target-texts with respect to target-language norms;

• between parallel subcorpora of corresponding pairs of source- and target-texts, to assess the amount of interlinguistic influence.

The corpus and methodology used in the present study attempt to expand on the bidirectional approach in several ways:

• Whereas results are often expressed as averages for the entire corpus, the text-by-text approach reveals the amount of variation within each subcorpus.

• Sentence-level alignment provides a means to overcome the limitations of macroscopic quantitative analysis through fine-grained qualitative analysis.

• The integration of neural machine translations (NMT) adds a new level of comparison between NMT and human translators.

This study will seek to determine to what extent human or machine-produced target-texts deviate from target-language norms, how much influence can be attributed to the source-texts, and whether human and machine-produced translations can be distinguished as separate subspecies using progressive forms (be+V-ing) in English and the periphrase “en train de+V-inf” in French as a basis of comparison.

2. Corpus and methods

2.1. Corpus preparation

A total of 35 public-domain works from the 19th-20th c. by 35 different authors in each language and the same number of translations (4×35 authors/translators = 140) were compiled into a ±13.5-million-word corpus consisting of four ±3.5m-word subcorpora. Source- and target-texts were aligned at sentence level and double-proofread. Machine translations were produced with DeepL Pro, bringing the total to ±20m-words with two more ±3.5m-word subcorpora. Finally, all texts were tagged in TreeTagger.

2.2. Data collection and analysis

Data were collected using regular expressions in AntConc. Results for each text were converted to normalized frequencies per 10,000 words (f/10k). Comparisons between subcorpora of target-texts and original texts representing the target-language were performed using the Wilcox-Mann-Whitney test, with Cohen's d as a measurement of effect-size. The correlation between source/target texts was evaluated using Spearman's correlation coefficient. Statistical analyses were carried out in R.

3. Results

3.1. Original English (En0) vs. Original French (Fr0)

Progressive forms are over 120 times more frequent in Original English (median 31.5/10k) than “en train de” in Original French (median 0.26/10k). The disparity is such that the respective frequencies cannot even be shown on the same scale:

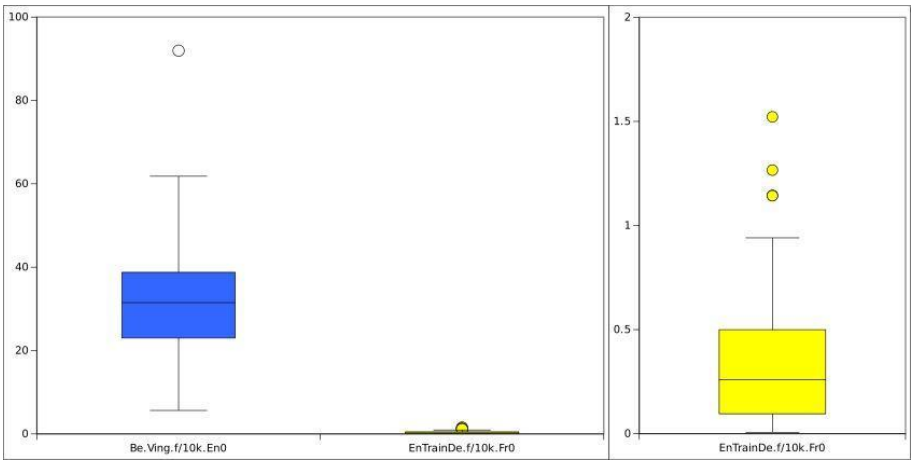


Fig. 1 Progressive forms in Original English (En0) and “en train de” in Original French (Fr0). Frequency of “en train de” in Fr0 (scale adjusted)

Paradoxically, however, translators have a tendency to overuse progressive forms in translations from French to English:

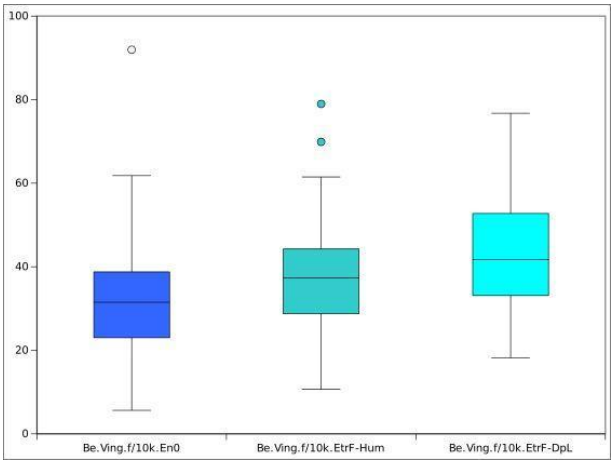


Fig. 2 Progressive forms in Original English (En0), English-translated-from-French by Human translators (EtrF-Hum) and by DeepL (EtrF-DpL)

3.3. Original French (Fr0) vs. French-translated-from-English (FtrE)

In translation from English to French, both human translators and DeepL use "en train de" much more often than it naturally occurs in Fr0 (median 0.26 in Fr0 vs. median 0.66/10k for FtrE-Hum, median 1.34/10k for FtrE-DpL, i.e. 2.5× and 5× respectively):

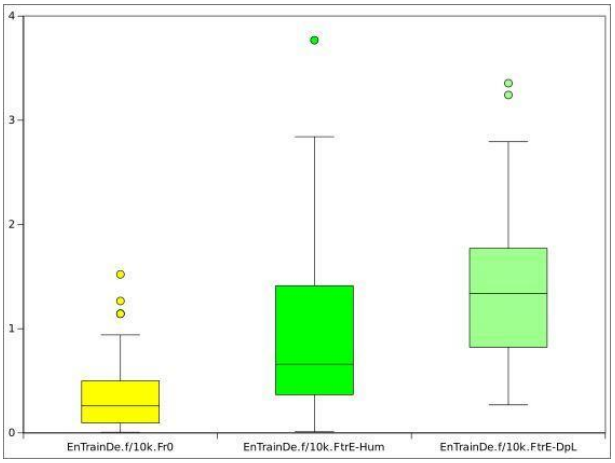


Fig. 3 "En train de" in Original French (Fr0), French-translated-from-English by human translators (FtrE-Hum) and by DeepL

Not only are both subcorpora distinct from Fr0, the range of frequencies observed in texts translated by DeepL sets it apart from human translators as well (p=0.01).

3.4. Source/Target-texts

Although the huge disparity between English and French, including French-translated-from-English, entails that only a very small proportion of progressive forms from En0 could potentially correspond to "en train de" in FtrE, a statistically significant correlation can nonetheless be found between the occurrences of be+V-ing in En0 source-texts and "en train de" in the corresponding FtrE target-texts. The correlation is noticeably stronger for DeepL (rho=0.74 vs. 0.46).

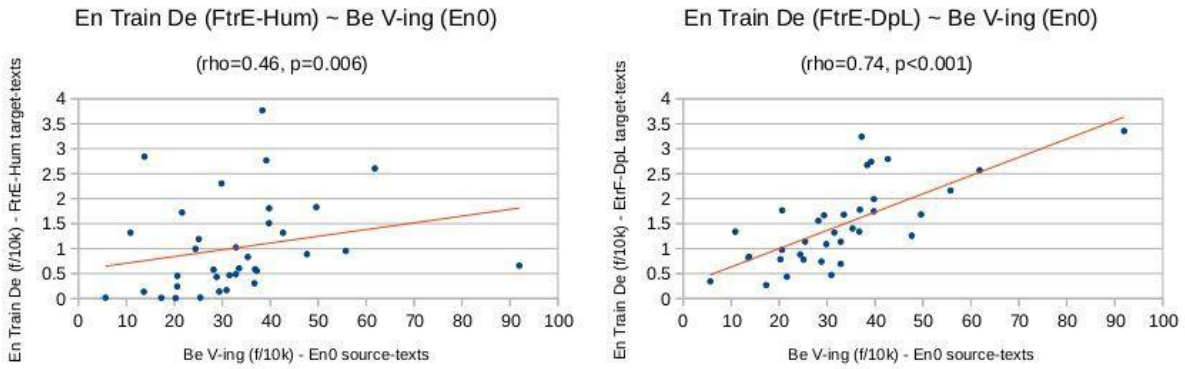


Fig. 4 Fig. Frequency of be+V-ing and "en train de" in corresponding source/target pairs: Human vs. En0 (left), DeepL vs. En0 (right).

In the other direction, no correlation can be detected between the tiny number of occurrences of "en train de" in Fr0 source-texts and the much higher frequency of be+V-ing in EtrF.

4. Discussion

The French periphrase "en train de" is often presented or perceived as equivalent or at least comparable in meaning to the progressive forms in English. Quantitative analysis demonstrates that, in terms of actual usage, the two are incommensurable with one another. Nonetheless, the perceived correspondence clearly has an impact on translation, both human and automatic.

Before the analysis can be carried any further, it must be emphasized that quantitative data are only suggestive of underlying differences. If human translators and DeepL use the present progressive more often than is statistically probable, it seems implausible that they use it in exactly the same ways. Overuse is more likely linked to specific usage patterns. This, however, must be verified through manual qualitative analysis. Manual analysis of random samples is presently underway and will provide a comprehensive account of when translators use these forms, what their counterparts are in the source-texts, and therefore how translated texts differ with respect to target-language norms.

References

Baker, M. (1993). Corpus Linguistics and Translation Studies. Implications and Applications. In *Text and Technology*, M. Baker, G. Francis, G. and E. Tognini-Bonelli (eds.). Benjamins: Amsterdam & Philadelphia, 233–250.

Frankenberg-Garcia, A. (2009). Compiling and using a parallel corpus for research in translation. *Babel: international journal of translation*, 21(1), 57-71.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In Lars Wollin & Hans Lindquist (eds.), *Translation Studies in Scandinavia*, 88–95. Lund: CWK Gleerup.

Granger, S., Lefer, M.-A. (2022). Corpus-based translation and interpreting studies: A forward-looking review. In S. Granger & M.-A. Lefer (eds), *Extending the Scope of Corpus-based Translation Studies*. Bloomsbury Advances in Translation series. London: Bloomsbury, 13-41.

Leech, G. (2007) New Resources, or just Better Old ones? The Holy Grail of Representativness. In: Hundt, Marianne/Nesselhauf, Nadja /Biewer, Carolin (eds) *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi, 133-149.

McEnery, T., Xiao, R. (2007). Parallel and comparable corpora. *Corpus-based perspectives in linguistics*, 131-145.

Teich, E. (2003). Cross-linguistic variation in system and text: a methodology for the investigation of translations and comparable texts. Berlin: Mouton de Gruyter.

Zanettin, F. (2011). Translation and corpus design. SYNAPS – *A Journal of Professional Communication* 26/2011, 14-23.

Software

Anthony, L. (2020). AntConc (Version 3.5.9). Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Farkas, A. (2012). LF Aligner (Version 3.11 Linux). https://sourceforge.net/projects/aligner/

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Schmid, H. (1994-2021). TreeTagger, Universität Stuttgart, http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

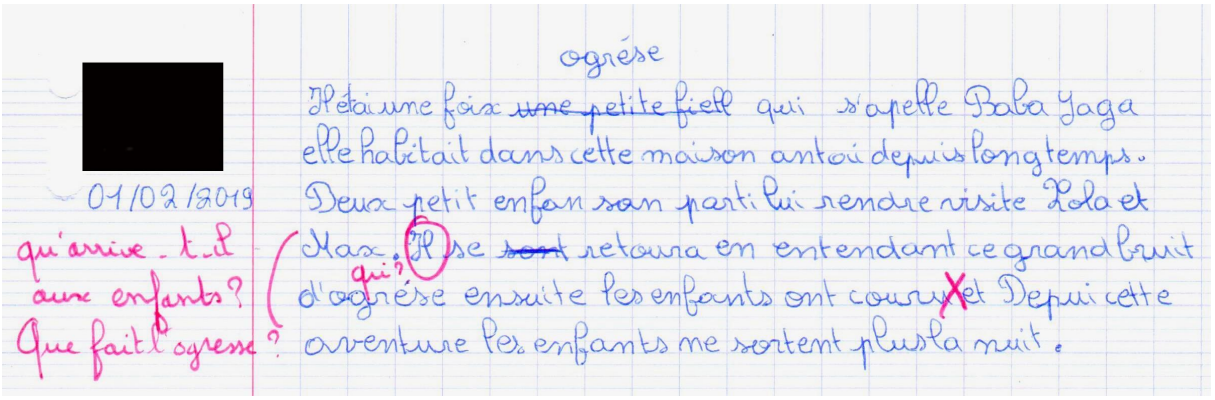
Un corpus de référence pour l'écriture de l'école à l'université : la ressource É-Calm

Doquet Claire [1], Maï Ho-Dac [2] et Claude Ponton [3]
[1] Laboratoire LabE3D, INSPE, Université de Bordeaux
[2] Laboratoire CLLE, Université de Toulouse Jean Jaurès
[2] Laboratoire LIDILEM, Université Grenoble Alpes
claire.doquet@u-bordeaux.fr, lydia-mai.ho-dac@univ-tlse2.fr, claude.ponton@univ-grenoble-alpes.fr

Introduction

Les écrits d'élèves intéressent depuis longtemps les linguistes, et singulièrement aujourd'hui, la linguistique de corpus [Doquet *et al*., 2017a]. Ils font partie des écrits non standards qui posent à l'analyse automatique de redoutables problèmes mais dont l'intérêt est évident pour travailler sur l'écriture manuscrite, la production spontanée de texte et l'utilisation de l'écrit chez des scripteurs non experts [Steuckardt et Collette, 2019]. Recueillis tout au long de la scolarité, ces ensembles de données tracent des trajets développementaux de l'acquisition de la langue écrite et de ses usages. La mise au jour de ces trajets est pour la linguistique un défi technologique et théorique. Technologique, parce que l'outillage de la linguistique de corpus ayant été pensé pour des textes proches des normes de l'écrit, les écrits scolaires mettent à l'épreuve les approches et les ressources classiques du traitement automatique des langues. Théorique, parce que la créativité langagière des élèves va au-delà de ce que prévoient les modèles élaborés sur des écrits standards, obligeant parfois à reconsidérer des catégories qui semblaient aller de soi. La situation d'apprentissage de l'écriture et les difficultés qu'elle révèle permettent en effet de mettre au jour les zones les plus résistantes de la langue qui apparaissent aussi, mais sous forme atténuée, chez les scripteurs disposant d'un haut degré de maîtrise de l'écrit.

À partir d'un corpus d'écrits d'élèves et d'étudiants rendu accessible sur une plateforme dédiée, le projet É–Calm (Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactique), financé par l'ANR, a permis de caractériser certaines compétences scripturales (orthographe et cohérence textuelle) et de mieux comprendre la manière dont les enseignants, par leurs interventions sur les copies, orientent l'écriture, afin d'étayer l'accompagnement de la réécriture de l'école à l'université. L'objectif principal de la recherche était de mettre à disposition en *open access* un grand nombre d'écrits produits dans des contextes d'apprentissage variés afin d'en analyser les caractéristiques linguistiques et discursives. L'ensemble ainsi constitué donne une visibilité aux textes de scripteurs à différents niveaux d'apprentissage et de maîtrise de l'écrit. La figure 1 donne un exemple de texte composant ce corpus.



Exemple de production d'un enfant de CM1

L'enjeu scientifique et sociétal d'une telle ressource est d'objectiver le regard sur les écrits scolaires. Les spécificités de ces écrits ont guidé les traitements et les analyses du corpus :

- Les écrits scolaires s'inscrivent dans une situation de communication particulière : celle d'un apprenant répondant à une commande de son enseignant, dans un contexte didactique donné (niveau de classe, consigne, environnement de la tâche...). L'organisation du corpus et l'accès aux écrits sont guidés par ces métadonnées contextuelles.
- Les écrits scolaires portent la trace de leur élaboration par l'élève auteur : le choix est fait de reprendre, pour la transcription de ces écrits, les catégories de la génétique textuelle. Elles permettent de mettre l'accent sur la dynamique de l'écriture et pas seulement sur les écrits terminés, elles donnent à lire, à travers les écrits des élèves, le trajet de l'écriture, outillant une didactique de l'écriture au sens plein du terme (comme l'hésitation entre parler d'une petite fille ou d'une ogresse dans la figure 1).

- Les écrits scolaires portent la trace des réactions de l'enseignant lecteur et correcteur, qui portent sur au moins deux dimensions de l'écrit : l'orthographe et la cohérence/cohésion textuelle (voir les interventions en rose dans la figure 1). É–Calm propose une typologie des interventions enseignantes, y compris non verbales (soulignements), ainsi qu'une catégorisation des corrections orthographiques et des marques de cohérence - ou d'incohérence - repérées sur les copies.

Le corpus É–Calm

La ressource É–Calm compte aujourd'hui près de 4500 textes (pour plus d'un million de mots) recueillis entre le début de l'école élémentaire et l'université et mis à disposition sur Ortolang (https://www.ortolang.fr/market/corpora/e-calm) et une plateforme dédiée sur HumaNum (http:// e-calm.huma-num.fr/). Conçue à la fois comme une vitrine et un lieu de diffusion des résultats mais aussi des regards que des scientifiques peuvent porter sur les écrits des élèves, cette plateforme s'adresse aux chercheurs mais aussi au monde enseignant, notamment les formateurs.

La constitution de corpus scolaires numériques est relativement récente. Le premier de ce type est sans doute le corpus Lancaster [Smith *et al.*, 1998] qui propose des textes d'enfants manuscrits retranscrits et accessibles en ligne. Toujours en anglais, l'Oxford Children's Corpus [Banerji *et al.*, 2013] propose depuis 2006 plus de 70.000 textes courts écrits par des enfants âgés de 4 à 13 ans dans le cadre de concours publics de rédaction en ligne. En 2011,

l'université de Karlsruhe diffuse un corpus de textes spontanés d'enfants allemands du grade 1 à 8 [Lavalley *et al.*, 2015]. Le premier corpus scolaire français remonte à 2005 et comporte 500 textes écrits d'enfants de CM2 à 5ème [Elalouf, 2005]. Depuis 2010, plusieurs projets de corpus scolaires sont en cours ou achevés [Garcia-Debanc et Bonnemaison, 2014, Doquet *et al.*, 2017b, Boré et Elalouf, 2017, Vogüé *et al.*, 2017, Wolfarth *et al.*, 2017] dont certains sont à l'origine du corpus É–Calm.

Les textes composants le corpus É–Calm sont issus de 4 projets pré-existants qui montrent certaines différences en fonction (a) de la consigne d'écriture qui, dans certains cas, a été proposée par les chercheurs du projet; (b) du niveau d'étude considéré; (c) de l'intervention ou non de l'enseignant sur la copie avec possibilité de réécriture.

Les corpus *EcriScol*[54] [Doquet *et al.*, 2017b] et *Littératie Avancée*[55] [Jacques et Rinck, 2017] sont composés de textes d'élèves et d'étudiants produits en réponse à une demande de leurs enseignants. Une grande partie de ces textes contient des interventions de ces enseignants et parfois même, pour *Ecriscol*, les brouillons et les versions intermédiaires. à noter que le corpus *Littératie Avancée* est le seul ne contenant que des textes informatisés. Les corpus *ResolCo*[56] [Garcia-Debanc *et al.*, 2017] et *Scoledit*[57] [Wolfarth *et al.*, 2017] ne comportent que des textes provoqués par la recherche.

Le corpus *ResolCo* se caractérise par une consigne élaborée pour confronter l'élève à une "tâche problème" nécessitant la gestion de liens de cohésion e.g. relations coréférentielles, relations de discours, etc. pour construire un texte cohérent. Cette "tâche-problème" consiste à rédiger une histoire incluant 3 phrases prédéfinies incluant, entres autres, des anaphores, des temps verbaux spécifiques et un enchaînement d'événements impliquant certaines relations de discours.

Scoledit est un corpus longitudinal permettant l'étude de l'évolution des compétences en littéracie durant l'école primaire [Wolfarth *et al.*, 2018]. Ce corpus est composé de textes produits selon les mêmes consignes par les mêmes 373 élèves durant toute leur scolarité primaire. Des productions narratives et des dictées ont ainsi été recueillies du CP au CM2 entre 2014 et 2018.

La table 1 fournit un aperçu quantitatif du nombre de textes de la version actuelle du corpus É–Calm.

table 6. : Niveau	**table 7. : #Textes**	**table 8. : #Mots**	**table 9. : #Mots/Textes**	**table 10. : Corpus**
table 11. : Primaire	table 12. : 133	table 13. : 30 214	table 14. : 105	table 15. : [E][R][S]
table 16. : Secondaire	table 17. : 55	table 18. : 15 866	table 19. : 209	table 20. : [E][R]
table 21. : Supérieur	table 22. : 07	table 23. : 71 056	table 24. : 110	table 25. : [E][R][LA]
table 26. : Total	**table 27. : 295**	**table 28. : 117 136**	**table 29. : 260**	**table 30. : [E][R][S][LA]**

[54] http://www.univ-paris3.fr/ecriscol
[55] https://www.ortolang.fr/market/corpora/litteracieavancee
[56] http://redac.univ-tlse2.fr/corpus/resolco.html
[57] http://www.scoledit.org/scoledit

table 31. : Etat quantitatif de la version actuelle du corpus É–Calm, avec [LA] pour Littéracie avancée, [E] pour Ecriscol, [R] pour Resolco et [S] pour Scoledit

Traitements manuels et automatiques des données

Tous les textes ont été récoltés sous la forme de copies manuscrites (ou tapuscrites pour ceux récoltés à l'Université) avec autorisation de diffusion. En plus de l'étape de transcription requis pour les copies manuscrites, les données ont nécessité un long processus d'encodage pour permettre une homogénéisation et une structuration des données et des méta-données associées. La norme TEI-P5 a été appliquée afin d'assurer le partage entre chercheurs de différentes disciplines, la pérennité des données et la possibilité d'appliquer directement des traitements automatiques prêts à l'emploi pour les corpus encodés selon la TEI. Cette norme a également été très précieuse pour guider l'encodage des métadonnées et des aspects génétiques comme les traces de révision (rature, ajout...)[58].

Une fois l'encodage des données brutes au format TEI-P5, l'étape suivante a consisté à proposer une normalisation de l'orthographe et à aligner les transcriptions et les versions normalisées. Cette étape a été réalisée manuellement selon deux méthodes selon les ressources : par la création en parallèle d'une version normalisée (comme en traduction) pour *Scoledit* ou via une interface d'annotation e.g. Glozz[59] [Mathet et Widlöcher, 2009], considérant alors la normalisation orthographique comme une première couche d'annotation. La première méthode requiert une étape d'alignement supplémentaire qui a nécessité le développement d'un outil spécifique : AliScol [Wolfarth *et al.*, 2018].

La mise à disposition des copies transcrites alignées à leur version normalisée permet l'application de traitements automatiques pour enrichir les données par l'annotation des lemmes, des étiquettes morphosyntaxiques et des relations de dépendances syntaxiques. L'ensemble de ces annotations a permis un ensemble d'analyses caractérisant les erreurs observées dans la ressource [Lavieu-Gwozdz *et al.*, 2021, Ponton *et al.*, 2021, Wolfarth *et al.*, 2018].

Un dernier pan de traitements manuels a consisté à annoter différents aspects de l'organisation discursive des textes d'élèves. Ainsi, un extrait du corpus *Resolco* a été segmenté en Unités Minimales de discours et annoté en relations de discours permettant d'éprouver les modèles du discours à ce type de données [Bras, 2021]. Enfin, près du quart de la ressource a été annotée en continuités référentielles, ce qui permet d'analyser en détail les stratégies mises en œuvre par les élèves pour tisser des liens référentiels dans leurs narrations [Garcia-Debanc, 2021].

La communication permettra de présenter la ressource É–Calm et les résultats de ces nombreuses analyses.

[58] . Voir le tableau 2 en annexe.

[59] . http://glozz.free.fr/

Références bibliographiques

Banerji, N., Gupta, V., Kilgarriff, A., Tugwell, D. (2013). Oxford children's corpus : A corpus of children's writing, reading and education. *Corpus Linguistics*, pages 315–317.

Lavieu-Gwozdz, B., Vinel, E., Goossens, V., Brissaud, C. (2021). Cartographie des usages et des erreurs orthographiques sur les verbes dans des récits écrits par des élèves de 6 à 15 ans. *Langue Française*, 211(3):51–65.

Boré, C., Elalouf, M.-L. (2017). Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles. *Corpus*, 16:31–64.

Wolfarth, C., Ponton, C., Brissaud, C. (2018). Gestion de la morphographie verbale en production d'écrits : que peut nous apprendre un corpus longitudinal? *Repères* [En ligne], 57.

Ponton, C., Gutiérrez-Cáceres, R., Teruggi, L., Brissaud, C., Wolfarth, C. (2021). Scolinter : un corpus trilingue. L'exemple de la segmentation en mots. *Langue Française*, 211(3):37–50.

Garcia-Debanc, C., Rebeyrolle, J., Ho-Dac, L.-M. (2021). La continuité référentielle dans le corpus rÉsolco : Méthode d'annotation et premières analyses. *Langue Française*, 211(3):99–114.

Doquet, C., David, J., Fleury, S. (Eds) (2017a). Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement. *Corpus*, 16 (Special Issue). OpenEdition.

Doquet, C., Enoiu, V., Fleury, S., Maziotti, S. (2017b). Problèmes posés par la transcription et l'annotation d'écrits d'élèves. Corpus, 16:133–156.

Elalouf, M.-L. (2005). *Ecrire entre 10 et 14 ans : Un corpus, des analyses, des repères pour la formation*. Canopé - CRDP de Versailles.

Garcia-Debanc, C., Bonnemaison, K. (2014). La gestion de la cohèsion textuelle par des élèves de 11-12 ans : Réussites et difficultés. In *Actes du 4e Congrès Mondial de Linguistique Française*, 961–976.

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, 16.

Jacques, M.-P., Rinck, F. (2017). Un corpus de littéracie avancée : résultat et point de départ. *Corpus*, 16.

Lavalley, R., Berkling, K., Stücker, S. (2015). Preparing children's writing database for automated processing. *In Proceedings of LTLT@SLaTE*, 9–15.

Mathet, Y., Widlöcher, A. (2009). La plate–forme GLOZZ : environnement d'annotation et d'exploration de corpus. In *Actes de TALN 2009*, Selnis. ATALA, LIPN.

Bras, M., Vieu, L., *et al*. (2021). Vers un corpus de textes d'élèves annoté en relations de discours. *Langue Française*, 211(3):115–129.

Smith, N., McEnery, T., Ivanic, R. (1998). Issues in transcribing a corpus of children's handwritten projets. *Literacy and Linguistic Computing*, 13:217–225.

Steuckardt, A., Collette, K. (2019). *Écrits hors-normes.* Les Éditions de l'Université de Sherbrooke (ÉDUS).

Vogüé, D., S. Espinoza, N., Garcia, B. ad Perini, M., Marzena Watorek, F. (2017). Constitution d'un grand corpus d'écrits émergents et novices : Principes et méthodes. *Corpus*, 16:65–86.

Wolfarth, C., Ponton, C., Brissaud, C. (2018). Gestion de la morphologie verbale en production d'écrits : que peut nous apprendre un corpus longitudinal ? *Repères*, 57.

Wolfarth, C., Ponton, C., Totereau, C. (2017). Apports du tal à la constitution et à l'exploitation d'un corpus scolaire. *Corpus*, 16.

The Structural Position Points Toward Different Functions: The Case of *For Sure*

Erina Iwai

Center for General Education, Shinshu University

Introduction

In present-day colloquial English, *for sure* expresses epistemic modality, that is, a speaker's judgement about the certainty of a clause (proposition) (cf. Biber et al., 1999, p. 854; Quirk et al., 1985, p. 620). (1) shows examples extracted from the *Corpus of Contemporary American English* (COCA).

(1)

a **For sure**, there are huge parallels. (COCA, SPOK, 2012)

b But eventually, yeah, they will, **for sure**. (COCA, SPOK, 2013)

c Velshi: Do you think you got a good mentor?
Gardenhour: **For sure**. (COCA, SPOK, 2005)

In (1a, b), *for sure* appears in the periphery[60]: the left periphery (LP, or initial position) in (1a) and the right periphery (RP, or final position) in (1b). It can also occur as an isolated response that encodes a positive polarity (affirmation), as in (1c). Expressing a speaker's epistemic stance of subjectivity, *for sure* serves as a pragmatic marker (PM; cf. Brinton, 2017).[61]

The study of periphery began amidst the development of linguistic fields such as discourse studies, historical pragmatics, and grammaticalization. Periphery is "where pragmatic meaning is negotiated" (Onodera, 2017; cf. Ohori, 1998, p. 194), so many researchers have tried to figure out what functions PMs have at a slot (e.g., Beeching & Detges, 2014; Olmen & Šinkūnienė, 2021; Onodera, 2017). Yet, figuring this out has not been as easy as it might seem, since pragmatic, interactional meanings are an "implication" given relative to the meaning of a form in correlation with individual context (cf. Takiura, 2008, pp. 150–1, 154). Such "meaning aspects are unpredictable from the lexical items involved[;] the structural context may be best described as a grammatical construction, a form–meaning pair" (Fischer, 2010, pp. 199–200).

The purpose of this study is to examine *for sure*, incorporating interactional notions (turn, turn-taking, action, etc.) and show that its pragmatic meanings are positionally sensitive and

[60] This study defines "periphery" as "the site in initial or final position of a discourse unit where metatextual and/or metapragmatic constructions are favored and have scope over that unit" (Traugott, 2017, p. 63). A "discourse unit" is often an utterance or turn in speech.

[61] *For sure* may also be used within a clause as a subjunct (Quirk et al., 1985, pp. 583–9); e.g., *I didn't know if I would* for sure *make it out* (COCA, SPOK, 2019). This use is outside the scope of this study.

can be inferred and predicted in a structure while, at the same time, it has a semantic role of modality.

Corpus and Methodology

Corpus

The corpus used for this study is COCA, which is a vast and representative corpus of American English. It contains more than one billion words from 1990–2019, covering eight genres: spoken, fiction, magazines, newspapers, academic texts, TV/Movies, blogs, and other web pages.

Methodology

Although COCA is genre-balanced, this study has confined itself to a spoken genre as the PM *for sure* is a phenomenon predominantly found in spoken language. COCA was searched using punctuation (search strings: * *for sure* , and * *for sure* .) to make it easier to collect the target instances, that is, *for sure* at the periphery and the stand-alone *for sure*. Then, the author manually excluded data on non-targeted uses. In total, 577 examples were collected, each of which was examined in terms of speaker-hearer interaction.

Results

Figure 1 *summarizes* the findings of the analysis.

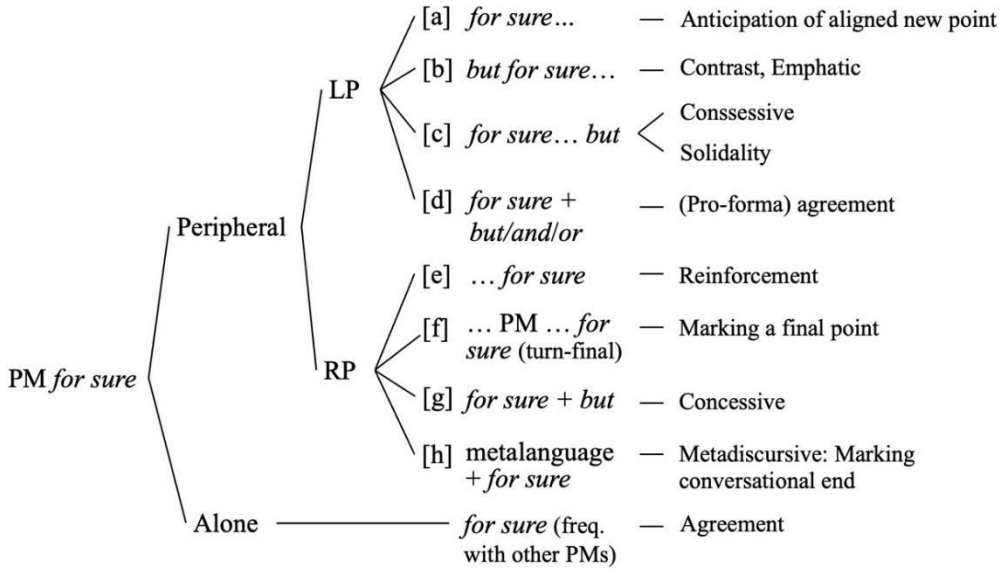


figure 1. : Schematization of the use of the PM *for sure*

First, *for sure* can be divided into peripheral and stand-alone use. The former is further divided according to LP or RP. At both peripheries, *for sure* exhibited a speaker's certainty about what he/she said, but what was notable was the fact that the RP *for sure* attached to units including evaluative adjectives more frequently than the LP *for sure*. (2) provides some examples.

(2)

a I credit it to a *great* crew, **for sure**. (COCA, SPOK, 2006)
b I just – I know he's going to have to be *tough*, **for sure**. (COCA, SPOK, 2010)
c That's *better* than bottled water, **for sure**. (COCA, SPOK, 2004)

Table 1 gives the number of occurrences of *for sure* with adjectives. The RP *for sure* co-occurs with evaluative adjectives four times more than the LP *for sure* (23.3% vs. 5.8%). Beltrama (2018) utilized evaluative adjectives (and other linguistic elements) as diagnostics to distinguish functions of the English *totally*, and found that *totally* co-occurs with evaluative adjectives (such as *brilliant*, *beautiful*, and *nice*) when it is used in a pragmatic meaning. Aijmer (2022), following Martin and White (2005), argues that evaluative adjectives are closely related to expression of positive and negative feelings. The result in Table 1 indicates that the PR *for sure* is likely associated with the speaker's expression of feelings, either positive or negative. In addition, it was found that epistemic modals are likelier to co-occur with the RP *for sure*, as seen in Table 1.

	Adjective (%)		Epistemic Modal (%)		
	Evaluative	Dimensional	Subjective	Objective	Both
Initial Total: 52	4 (7.7)		5 (9.6)		
	3 (5.8)	1 (1.9)	5 (9.6)	0	0
Final Total: 262	68 (26.0)		49 (18.7)		
	61 (23.3)	7 (2.7)	34 (13.0)	13 (5.0)	2 (0.8)
evaluative adjectives: e.g., *great*, *hard*, *dramatic*, etc.; dimensional adjectives: e.g., *short*, *slow*, *huge*, etc.; epistemic modals (subjective): e.g., *I know*, *would*, etc.; epistemic modals (objective): e.g., *probably*, *certainly*, etc. (cf. Martin and White, 2005, pp.13–4)					

table 1. : The number of occurrences of *for sure* with adjectives and epistemic modals

(3)

And *I think* that there are *absolutely* fair criticisms to be made of Jared and Ivanka's role in the Trump White House, **for sure**. (COCA, SPOK, 2019)

In (3), the RP *for sure* is used with other epistemic modals, *I think* and *absolutely*. Uttering *for sure* after these modals reinforces the speaker's commitment to the content of the utterance.

Next, multiple structures were observed related to the use of LP/RP *for sure*: [a]–[h] in Figure 1. In each structure, *for sure* has a particular pragmatic, interactional function (the rightmost column in Figure 1). Here, examples of [b], [c], [f], and [h] are given.[62]

(4)

b Neal-Conan: Shinjiro Murata, Deborah Amos has been telling us about shortages of food. Are you seeing malnutrition in your patients?
Shinjiro-Murata: Yes. Sometimes we have malnutrition children among under five years old but not – it's not majority yet. **But for sure**, we are receiving also a lot of report of shortage of food, bread, created by the shortage of fuel. (COCA, SPOK, 2013)

c Well, **for sure**, there are a lot of distractions that are facing today's drivers. **But** one of the things that we know is that all distractions are not equal. (COCA, SPOK, 2011)

f I did. Not at the very beginning, but I did eventually go to the morgue. **And**, **yeah**, it is – it's a – it's an experience you never forget, **for sure**. (COCA, SPOK, 2016)

h Well, thanks so much for sharing your story. **It's an important one**, **for sure**. (COCA, SPOK, 2019)

In [b], *but for sure* strengthens the contrast with the previous utterance. It is emphatic in that it expresses the speaker's strong agreement to what the interlocutor has said. In [c], *for sure* represents a concession by working with the postpositioned *but*. This *for sure…but*-construction may be considered a "projector construction" (Günthner, 2015; Shibasaki, 2014, 2021a). In [f], *for sure* accompanies the last utterance (with the LP *and* and *yeah*) of a turn and reconfirms the final point. In [h], *for sure* attaches to a metaconversational utterance, marking the end of the conversation.

Thus, pragmatic, interactional meanings of *for sure* are determined and predictable in the structure (the pattern of the information sequence) beyond sentences and clauses in individual contexts.

So far, constructions similar to *for sure*'s have been reported crosslinguistically: e.g., *zwar…aber* ('true…but') in German (Günthner, 2016) and *tashika ni…shikashi/demo* ('surely…but') in Japanese (Shibasaki, 2021b). It is believed that accumulating findings about constructions in interaction in the corpus data will shed further light on human interaction, more specifically, "thetical" grammar (Kaltenböck et al., 2011) pertain to the situation of the discourse (e.g., attitude of the speaker, text organization, speaker-hearer interaction).

Bibliographic references

Aijmer, K. (2022). "That Is Totally Not My Type of Film": Innovations in the Intensifier System of UK English. In E. Peterson, T. Hiltunen, J. Kern (Eds.), *Discourse-Pragmatic*

[62] In the presentation, the author will present examples with a broader context.

Variation and Change: Theory, Innovations, Contact (pp. 127–49). Cambridge: Cambridge University Press.

Beeching, K., Detges, U. (Eds.). (2014). *Discourse Functions at the Left and Right Periphery: Crosslinguistic Investigations of Language Use and Language Change*. Leiden/Boston: Brill.

Beltrama, A. (2018). *Totally* Between Subjectivity and Discourse: Exploring the Pragmatic Side of Intensification. *Journal of Semantics*, 35.2, 219–61.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Brinton, L. J. (2017). *The Evolution of Pragmatic Markers in English: Pathways of Change*. Cambridge: Cambridge University Press.

Fischer, K. (2010). Beyond the Sentence: Constructions, Frames, and Spoken Interaction. *Constructions and Frames* 2.2, 185–207.

Günthner, S. (2015). A Temporally Oriented Perspective on Connectors in Interactions: *Und Zwar* ('Namely/In Fact')-Constructions in Everyday German Conversations. In A. Deppermann, S. Günthner (Eds.), *Temporality in Interaction* (pp. 237–64). Amsterdam: John Benjamins.

Günthner, S. (2016). Concessive Patterns in Interaction: Uses of *Zwar…Aber* ('True…But')-Constructions in Everyday Spoken German. *Language Sciences* 58, 144–62.

Kaltenböck, G., Heine, B., Kuteva, T. (2011). On Thetical Grammar. *Studies in Language* 35, 852–97.

Martin, J., White, P. R. (2005). *The Language of Evaluation. Appraisal in English*. Basingstoke: Palgrave Macmillan.

Ohori, T. (1998). Close to the Edge: A Commentary on Horie's Paper. In T. Ohori (Ed.), *Studies in Japanese Grammaticalization: Cognitive and Discourse Perspectives* (pp. 193–7). Tokyo: Kurosio Publishers.

Olmen, D. V., Šinkūnienė, J. (2021). *Pragmatic Markers and Peripheries*. Amsterdam/ Philadelphia: John Benjamins.

Onodera, N. (Ed.). (2017). *Hatsuwa no Hajime to Owari—Goyoronteki Chosetsu no Nasareru Basho* (*Periphery: Where pragmatic Meaning is Negotiated*). Tokyo: Hituzi Syobo.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Shibasaki, R. (2014). On the Grammaticalization of *The Thing Is* and Related Issues in the History of American English. In M. Adams, R. D. Fulk, L. J. Brinton (Eds.), *Studies in the History of the English Language: Evidence and Method in Histories of English* (pp. 99–122). Berlin: De Gruyter Mouton.

Shibasaki, R. (2021a). Reanalysis and the Emergence of Adverbial Connectors in the History of Japanese. In A. Haselow, S. Hancil (Eds.), *Studies at the Grammar-Discourse Interface:*

Discourse Markers and Discourse-Related Grammatical Phenomena (pp. 101–24). Amsterdam: John Benjamins.

Shibasaki, R. (2021b). *Tashika Ni* no Danwakino to Teikeisei ni Tsuite. Paper presented at *Online International Symposium "Formulaicity in Interactional Discourse,"* March 2021.

Takiura, M. (2008). *Poraitonesu Nyumon* (*Politeness*). Tokyo: Kenkyusha.

Traugott, E. C. (2017). A Constructional Exploration into "Clausal Periphery" and the Pragmatic Markers That Occur There. In N. Onodera (Ed.), *Hatsuwa no Hajime to Owari—Goyoronteki Chosetsu no Nasareru Basho* (*Periphery: Where pragmatic Meaning is Negotiated*) (pp. 55–73). Tokyo: Hituzi Syobo.

Corpus

The Corpus of Contemporary American English 1990–2019 (COCA). (M. Davies), available online at https://www.english-corpora.org/coca/.

Approche du français *de tous les jours* en classe de FLE à la lumière d'un corpus de messages vocaux

Laure Anne Johnsen
[1] ILCF, Université de Neuchâtel, Suisse
laure.johnsen@unine.ch

Introduction

Dans cette communication, nous relatons une expérience d'enseignement dans le cadre d'un cours universitaire pour allophones consacré à l'apprentissage français du quotidien. Cette expérience s'inscrit dans le contexte du développement didactique du corpus OFROM (corpus oral de français de Suisse romande, Avanzi et al. 2012-2022, http://ofrom.unine.ch) en cours.

La *langue de tous les jours* (Blanche-Benveniste 1985, 1997), en Suisse romande comme ailleurs, est rarement prise en compte dans l'enseignement du français. En effet, lorsqu'on parcourt les manuels de français langue de scolarisation ou étrangère, l'oral du quotidien fait figure de parent pauvre à côté du français dit « de référence », lequel repose fondamentalement sur une norme écrite même lorsqu'il s'agit d'exercer l'oral (Gagnon & Benzitoun 2020, Surcouf & Ausoni 2022). Or, l'écart entre le français appris dans un cadre institutionnel et celui réellement pratiqué par les natifs est régulièrement rendu explicite, voire déploré, par des apprenants se retrouvant subitement en immersion dans une région francophone (Molnár 1999 : 17, Durán et McCool 2003, Weber 2006, French et al. 2017 : vi, Paternostro 2016 :123 , cités par Surcouf et Ausoni 2022 :133).

Malgré l'existence de quatre langues nationales et les efforts d'une politique éducative ouverte au plurilinguisme et aux échanges linguistiques, la Suisse n'est pas épargnée par ce constat, comme l'illustrent ces témoignages d'étudiantes sur leur parcours d'apprentissage respectif du français à l'école en Suisse alémanique et en Suisse italienne :

1) Il me manquait de vocabulaire pour des situations spécifiques. En outre, j'avais quelques **problèmes de comprendre le français parlé dans un contexte hors de l'école**. » (étudiante originaire de Saint-Gall, 2022)
2) [à propos d'un séjour dans une famille d'accueil francophone] « C'est surprenant qu'**après cinq ans de français je n'étais pas capable de vraiment parler. Mais c'est vrai que si on parle à l'école, souvent ce sont des présentations appris, sur des thématiques dont on parle rarement dans un vrai dialogue**. J'ai certainement appris quelque chose au niveau de la langue, mais l'aspect le plus important était que l'intérêt d'une langue ne doit pas être de connaitre le plus de mots possibles par cœur, ou de connaitre la grammaire, mais de **parler, de discuter et de savoir utiliser cette langue**. » (étudiante originaire de Bâle, 2021)
3) Si je considère encore aujourd'hui ces leçons, et en général la façon dont les langues sont enseignées au Tessin, comme inadéquates pour un vrai apprentissage, c'est pour des raisons précises. Bien que la grammaire soit fondamentale pour comprendre les règles d'une langue comme les modes et temps verbaux et les pronoms, je suis de

> l'avis que **seulement en pratiquant et en écoutant cette langue on peut bien l'apprendre et être capable de l'utiliser dans le monde réel**. […] j'étais quand même soumise à une épreuve écrite mensuelle composée de plusieurs exercices de grammaire et du **vocabulaire que je devais apprendre par cœur, mais je n'aurais su ni prononcer correctement tous ces mots ni les utiliser dans un contexte donné dans la vie**. (étudiante originaire du Tessin, 2022)

À travers ces témoignages, la distance entre le contexte scolaire d'une part (« présentations appris » (sic), « règles d'une langue », « par cœur », « contextes plutôt artificiels ») et la pratique ordinaire de la langue d'autre part (« vrai dialogue », « monde réel », « contexte donné dans la vie », « situations réelles ») est systématiquement pointée du doigt. C'est donc dans l'objectif de fournir des moyens aux enseignants pour mieux préparer les apprenants au français parlé dans des situations du quotidien que le projet *Enseignement* d'OFROM a vu le jour, qu'il s'agisse de favoriser la cohabitation des citoyens, la mobilité professionnelle ou estudiantine, l'accueil de personnes de l'étranger pour des séjours occasionnels ou de longue durée ou simplement l'observation (depuis l'étranger) du français tel qu'il est parlé dans divers contextes en Suisse.

Les différents axes du développement d'OFROM seront brièvement exposés, à savoir a) la diversification des genres de parole du corpus (conversation, réunion, table ronde, conte, enseignement, vocaux, entretien d'embauche, etc.), b) les ressources à vocation documentaire/illustrative sur le français régional de Suisse et le français parlé en contexte (FAQ), c) les activités et outils destinés plus spécifiquement à l'enseignement du FLE. C'est dans le cadre de ce dernier axe que nous proposons de relater une expérience d'enseignement.

Méthodologie

L'expérience d'enseignement s'est tenue dans le cadre d'un cours universitaire consacré à l'oral du quotidien et destiné à un groupe hétérogène d'apprenants de niveau B1-B2. Les objectifs sont, parmi d'autres, l'amélioration de la compréhension et de l'expression à l'oral, la reconnaissance des caractéristiques du français parlé au quotidien et le développement de la *compétence sociolinguistique*, c'est-à-dire, la capacité de s'adapter à différents types d'interactions sociales (André & Tyne 2012).

A cet effet, nous avons élaboré une séquence didactique autour d'une pratique émergente de l'oral, à savoir le *message vocal* (appelé également *note vocale*, *audio* ou simplement *vocal*). Il s'agit de messages audio à caractère monologal, enregistrés et envoyés via des systèmes de messagerie (par ex. *Whats'app, Messenger, Signal*, etc.) à un destinataire unique ou collectif en mode asynchrone, relativement spontanés et qui relèvent en général de la proximité communicative (Glikman & Fauth 2022). Ce mode de communication nous a paru particulièrement adapté à une première approche de la langue de tous les jours, aux objectifs d'apprentissage du cours et au public d'apprenants concerné.

La première étape consistait en un travail d'observation en plenum d'un audio authentique et l'identification des faits remarquables à plusieurs niveaux d'analyse (phonétique, syntaxe, lexique, pragmatique/interaction). Par la suite et sur une période d'un mois, le travail demandé aux étudiants (une vingtaine) était le suivant : par groupe de deux ou trois, il s'agissait de recueillir un message vocal par personne (min. 30 sec.) auprès de locuteurs natifs de la région (Suisse romande). Les messages devaient être authentiques (non élicités) et déjà existants dans

la messagerie des informateurs au moment de la demande. A cela s'ajoutait le renseignement du contexte du message (relation avec le destinataire, date, etc.) et du profil socio-démographique du locuteur par les étudiants. À partir des messages récoltés, chaque groupe avait pour tâche l'écoute puis la sélection de l'un d'entre eux en vue d'une présentation devant la classe. Celle-ci consistait en une analyse du message et des faits d'oral remarquables selon les niveaux d'analyse respectifs (cf. *supra*) à partir d'une transcription orthographique préalable de l'audio, puis en une discussion des problèmes rencontrés au cours du travail.

Une deuxième phase de la séquence était consacrée à la production. Après l'observation de plusieurs audios illustratifs d'un acte de langage particulier (compliment, reproche, instructions, etc.) la mise en place de jeux de rôles impliquant des actes similaires ont permis le réinvestissement des éléments lexicaux, pragmatiques et interactionnels repérés. Enfin, la tâche finale consistait en l'enregistrement d'un audio : un message vocal (min. 30 sec.) devait être enregistré par chaque étudiant en réaction à un SMS donné (fictif), avec une attention particulière à porter sur l'adéquation au message (situation, contenu, actes de parole, lexique, proximité avec le destinataire, ménagement des faces, etc.).

Résultats et perspectives

Nous synthétiserons ainsi les apprentissages réalisés (observés par les apprenants et par les enseignants) et tirerons un bilan de cette expérience qui, dans la lignée d'autres travaux exploitant des corpus oraux dans une visée didactique[63], a pour ambition de confronter les apprenants à une image plus réaliste des usages variés de la langue cible.

Le récolte de ces données s'inscrit dans un projet plus vaste de constitution d'un corpus de vocaux en passe d'être intégré à la base OFROM, qui servira à mieux documenter, dans le sillage de Glikman & Fauth (2022), cette pratique de l'oral désormais répandue et à la mettre en perspective d'autres genres de discours. En outre, ce sous-corpus vise à poursuivre, autour des travaux d'Avanzi *et al.* (2016), l'étude des variétés de français parlées en Suisse romande.

Références bibliographiques

André, V. & Tyne, H. (2012). Compétence sociolinguistique et dysfluence en L2. In Kamber, A. & Skupien, C. (dir.). *Recherches récentes en FLE.* Berne : Peter Lang. 21-46.

Avanzi M., Béguelin M.-J., Corminboeuf G., Diémoz F. & Johnsen L. A. (2012-2023). Corpus OFROM – *Corpus oral de français de Suisse romande*. Université de Neuchâtel, www.unine.ch/ofrom

Avanzi M., Béguelin M.-J. & DIÉMOZ F. (2016). De l'archive de parole au corpus de référence : la base de données orales du français de Suisse romande (OFROM) », *Corpus*, 15. https://journals.openedition.org/corpus/3060

Blanche-Benveniste, C. (1985). La langue du dimanche. *Reflet*, 14, . 42-43.

[63] Parmi celles-ci, citons le projet PFC-EF https://www.projet-pfc.net/le-projet-pfc-ef/, la base CLAPI-FLE http://clapi.icar.cnrs.fr/FLE, la ressource FLORALE https://florale.unil.ch/ ou encore les travaux autour du site FLEURON https://fleuron.atilf.fr/. Les objectifs spécifiques, le public cible et les types de ressources peuvent toutefois sensiblement varier d'un projet à l'autre.

Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français*. Paris, Ophrys, France.

Gagnon, R., & Benzitoun, Ch. (2020). Le français parlé comme objet d'enseignement ? Regards croisés d'un linguistique et d'une didacticienne ». *Formation et Pratiques d'Enseignement en Questions - Revue des HEP de Suisse romande et du Tessin, 26*, 37-51

Glikman, J., Fauth, C. (2022). Un nouvel accès à la parole spontanée : les vocaux. *Proc. XXXIVe Journées d'Études sur la Parole* -- JEP 2022, 154-162.

Durán, R. & McCool, G. (2003) If This Is French, Then What Did I Learn in School ? *The French Review*, 77-2, 288-299.

French, L., Beaulieu S. & Huot, D. (2017). Regard sur le développement de la compétence de communication à l'oral :: récit rétrospectif d'un apprenant de français langue seconde. *Revue canadienne de linguistique appliquée*, 20-2, xxxiii.

Johnsen, L.A., (à par. 2023). Approche du français parlé dans une perspective d'enseignement : quelques pistes d'exploitation pédagogique du corpus OFROM, *Travaux de linguistique*.

Molnar, K. (1999), *Konférans pour lé zilétré*. Romainville : Al Dante.

Paternostro, R.(2016). *Diversité des accents et enseignement du français. Les parlers jeunes en région parisienne*. Paris, L'Harmattan.

Surcouf, Ch. & Ausoni (2022). "le français parlé ? Eh ben j'savais pas ce que c'était ! " : Production et compréhension de la variation diaphasique en français parlé en FLE. *Mélanges CRAPEL*, 43 (1), 130-156.

Weber, C., (2006). Pourquoi les Français ne parlent-ils pas comme je l'ai appris. *Le français dans le monde*. 345, Paris, CLE International, 31-33.

Constituer un corpus pour l'étude du *code-switching* dans la *Correspondance* de Cicéron

Cécile JULLION [1] et Julie SORBA [2]
[1] Laboratoire Litt&Arts-CNRS, Univ. Grenoble Alpes
[2] Laboratoire LIDILEM, Univ. Grenoble Alpes

Introduction

De nombreuses études considèrent la Correspondance de Cicéron comme « le plus vaste corpus de la littérature latine où se déploie le 'code-switching' » (Aubert-Baillot, 2021, p. 16), faisant du corpus épistolaire cicéronien un objet « authentique, c'est-à-dire composé de véritables lettres, envoyées à de véritables correspondants » (ibid.). En outre, ces lettres, qui « nous ouvrent la porte des coulisses de l'histoire, la 'grande histoire', celles des conflits politiques qui déchirent la République » (Robert et al., 2021, p. XI), présentent une occasion unique de saisir le concept de code-switching – défini comme « la juxtaposition, à l'intérieur d'un même échange verbal, de passages où le discours appartient à deux systèmes ou sous-systèmes grammaticaux différents » (Gumperz, 1989b, p. 57 : cité par Dabène, 1994, p. 93-94) – dans toute sa complexité et toute sa singularité, à une époque où le bilinguisme gréco-latin caractérise l'ensemble d'une société et en affecte toutes les classes.

Aussi, sachant qu'il existe une grande diversité d'emplois du code-switching chez les auteurs antiques, l'enjeu de cette contribution qui porte spécifiquement sur la Correspondance de Cicéron est de s'interroger sur la pluralité des typologies du code-switching pour choisir celle(s) qui nous permette(nt) de constituer un corpus cohérent, en adéquation avec l'étude syntaxique et sémantique des séquences grecques que nous menons. Pour ce faire, nous partirons des travaux effectués sur cet objet d'étude pour en exposer les principales limites puis nous détaillerons les différentes étapes qui nous ont conduit à établir de nouvelles typologies du grec, qui reflètent davantage la réalité du bilinguisme gréco-latin de la Rome républicaine de l'époque de Cicéron. Notre étude impliquera plusieurs champs de la linguistique, en fournissant aussi bien des données d'ordre syntaxique et sémantique que terminologique.

Corpus et méthodologie

Principes de constitution du corpus

L'authenticité est un paramètre central pour Firth qui considère que l'attestation dans des textes authentiques est essentielle à l'approche du sens : « An approach to the meaning of words, pieces, and sentences by the statement of characteristic collocations ensures that the isolate word or piece as such is attested in established texts » (Firth, 1957 [1951], p. XI). De plus, le corpus s'envisage sur le plan statistique comme un échantillon d'une population d'événements langagiers (Habert, 2000), que l'on souhaite représentatifs alors d'un

phénomène linguistique particulier : « However, it should also be remembered that the corpus itself is a sample and needs to be representative of a given aspects of language so that 'The first step towards achieving this aim is to define the whole of which the corpus is to be a sample' (Renouf, 1987, p. 2) » (Nelson, 2010, p. 56-57).

C'est précisément dans la perspective de l'échantillonnage de productions authentiques que nous inscrivons notre démarche : les lettres de la Correspondance sont de véritables lettres, adressées à des correspondants réels, des personnages historiques, tels que César ou Pompée. Notre échantillonnage opère au sein des sept premiers livres de la Correspondance de Cicéron. Le travail que nous présentons ici s'inscrit dans un cadre plus vaste qui sera complété avec le dépouillement des quatre derniers livres.

Parmi les séquences grecques de la Correspondance – qui ne compte pas moins de « 850 mots ou groupes de mots grecs dont aucun n'est employé à plus de trois reprises » (Aubert-Baillot, 2021, p. 16-17), nous avons choisi de présenter deux jeux de données en particulier :

- les séquences impliquant un ou plusieurs verbes y compris à l'infinitif et au participe – et, pour ces derniers, seuls ceux qui occupent une fonction verbale, car ce sont les deux parties du discours pour lesquelles Cicéron fait preuve du plus d'inventivité lexicale. Ce choix sera l'occasion d'analyser les néologismes (Pruvost et Sablayrolles, 2019), particulièrement instructifs sur l'oral d'une langue pour laquelle il ne nous reste que des traces écrites.
- les séquences contenant des citations grecques versifiées – définies comme des énoncés qui présentent une double contrainte de versification et de citation (Nicolas, 2006). Leur étude permet entre autres de questionner le concept de citation, notamment en détaillant la méthodologie qui nous a permis de les identifier en tant que telles et qui consiste non seulement en la confrontation des notes infra-paginales des différentes éditions de la Correspondance mais aussi au recours aux bases de données électroniques – essentiellement Perseus Digital Library et Code-Switching in Roman Literature – qui seront utilisées comme corpus de vérification.

Nous présentons dans le tableau ci-dessous de manière quantitative notre corpus d'étude, qui se compose donc des séquences grecques contenant des verbes et des citations grecques.

	Verbes (y compris infinitif)	Participes (avec fonction verbale)	Citations grecques versifiées	**TOTAL**
Livre 1	37	1	20	58
Livre 2	10	1	6	17
Livre 3	20	3	10	33
Livre 4	39 ± 2[64]	16 ± 2	10	65 ± 4
Livre 5	45 ± 1	20 ± 1	27	92 ± 2
Livre 6	12 ± 1	5	11	28 ± 1
Livre 7	10	2	4	16
TOTAL	173 ± 4	48 ± 3	88	**309 ± 7**

table 1. : **Présentation du corpus d'étude**

[64] Le signe ± indique les cas difficiles à désambiguïser qui nécessitent une analyse particulière, plus détaillée.

Notre corpus se compose donc de 309 ± 7 séquences grecques réparties comme suit : 173 ± 4 verbes, 48 ± 3 participes et 88 citations grecques versifiées. Sur un total de 850 mots ou groupes de mots grecs, cela constitue 36,35 % de l'ensemble du corpus, ce qui justifie pleinement notre choix. Par ailleurs, nous observons de fortes disparités entre les livres : le livre 2 en particulier, tout comme le livre 7, présentent des taux assez bas de verbes, de participes et de citations grecques versifiées. En fait, ceux-ci renferment des lettres rédigées à des périodes troublées de la vie de Cicéron, durant lesquelles il a dû faire face à des situations particulièrement difficiles, tant sur le plan affectif qu'émotionnel[65]. F

Or, le recours au grec par Cicéron dans ses lettres n'est jamais anodin[66] mais est fonction de plusieurs paramètres, à la fois internes et externes à l'environnement linguistique, tels que la recherche du « mot juste » ou le grec comme « langue de l'intimité » et du retour sur soi (facteurs internes) ou les divergences observées non seulement entre les différents correspondants mais aussi entre les diverses périodes de sa vie (facteurs externes).

Méthodologie pour établir une typologie des échantillons

Notre étude sur les séquences grecques dans la Correspondance de Cicéron s'inscrit dans la lignée des travaux de Steele (1900), Jackson (2014) et Adams (2003) dont les conclusions respectives sont présentées en annexes. Ceux-ci ont proposé deux types de typologies de ces séquences grecques : morpho-syntaxique pour Steele (1900) et Jackson (2014) et pragmatico-sémantique pour Adams (2003). Or, si les recherches menées par Steele et Jackson ont l'avantage de mettre en exergue le concept de citation pour le questionner, leur typologie – en particulier celle de Steele – reste très hétérogène et peu probante, notamment à cause de la dissociation faite entre le vers (catégorie intitulée « poets ») et la prose1 (catégorie intitulée « prose-writers »). Par ailleurs, nous pouvons aussi nous demander pourquoi la question des proverbes – qui comprend ce que Steele nomme « detached phrases » et « ciceronian phrases » (cf. annexes, tableau a) – n'est pas traitée sur le même plan que la dichotomie vers / prose adoptée pour l'étude des citations. En somme, leur corpus, qui consiste en une série de termes grecs classés en fonction des différentes parties du discours (« adverbs », « adjectives », « nouns » et « verbs » entre autres2), semble constitué sans véritable objectif d'analyse ou d'application pensé en amont. La démarche adoptée par Adams, qui propose une typologie3 pragmatico-sémantique du grec de la Correspondance, est certes pertinente en nous renseignant sur les possibles intentions de Cicéron à recourir au grec dans certaines lettres en particulier, mais la démarche onomasiologique qu'il choisit ne lui permet pas d'éviter les écueils de la surinterprétation et de l'interprétation psychologisante.

Notre approche, à la fois sémasiologique et onomasiologique, s'inspire des recherches menées par ces trois auteurs, mais en les nuançant et en les complétant. Progressant livre par livre, par le retour et la confrontation au texte, nous relevons à la fois les séquences grecques dans leur contexte – c'est-à-dire avec l'entour latin dans lequel elles s'insèrent – et de façon isolée. Pour la première typologie, d'ordre sémasiologique, le but est de dégager le caractère

[65] Les lettres du livre 2 ont été rédigées pendant la période de l'exil ; une partie de celles du livre 7 peu de temps après la mort de sa fille Tullia.

[66] Voir par exemple Rochette (2012, p. 94) qui fait l'hypothèse du grec comme « langue de l'intimité » et renvoie à « la vie psychologique et émotionnelle » du sujet parlant pour expliquer le recours au grec par Cicéron dans ses lettres. Pour ce faire, il se fonde entre autres sur le fait que César aurait prononcé ses derniers mots, avant de mourir, en grec : καὶ σὺ τέκνον... Sur les dernières paroles de César, voir l'article de Dubuisson M. (1980), « Toi aussi, mon fils ! », dans Latomus 39, p. 881-890 (spécialement les pages 888-890) et le lien qu'il fait entre les notions de « langue maternelle » et de « langue première ».

univoque d'une séquence grecque donnée, c'est-à-dire de la désambiguïser. La deuxième typologie est d'ordre plutôt sémantique, en classant les séquences grecques selon la terminologie des grands champs du savoir à l'époque antique (rhétorique, philosophie, médecine, etc.). Or, ces deux types de classement induisent inévitablement des obstacles : par exemple, pour les citations grecques versifiées, la pertinence du choix du critère du genre textuel soulève de nombreuses questions. Pour faire face à ces difficultés, et en particulier à la démarche terminologique qui sous-tend l'ensemble de la recherche, nous utilisons des corpus déjà constitués afin d'affiner le nôtre en nous fondant sur le présupposé suivant : le critère d'attestation d'un terme technique provient de l'attestation de ce terme dans un texte dans lequel il est reconnu avec sa valeur technique ; autrement dit, pour justifier la valeur technique d'un terme, nous nous appuyons à la fois sur la tradition de transmission des textes et sur les travaux antérieurs menés à ce sujet. Concernant les citations grecques versifiées, nous commencerons par fournir notre propre définition du terme citation ; et, si plusieurs critères sont envisageables pour classer les citations grecques de notre corpus, nous préférerons le critère métrique qui est plus objectif, notamment en termes de contraintes pesant sur l'énoncé. Pour appréhender le concept de citation, nous exploiterons les travaux menés par Compagnon (1979, p. 81) qui propose la définition suivante : « la citation est un énoncé répété et une énonciation répétante : en tant qu'énoncé, elle a un sens, 'l'idée' qu'elle exprime dans son occurrence première (t dans S1) ; en tant qu'énoncé répété, elle a également un sens, 'l'idée' qu'elle exprime dans son occurrence seconde (t dans S2) ». Notre définition de la citation renverra également à la notion de discours citationnel tel que Darbo-Peschanski (2004, p. 12) l'envisage : « la combinaison du discours rapporté et du discours d'accueil, un seul énoncé est attribué à deux locuteurs différents ».

Résultats

Si notre travail est encore en cours d'élaboration, nous pouvons néanmoins en présenter les premiers résultats et pistes d'interprétation, en particulier pour les citations grecques versifiées. En effet, nous prévoyons de dresser un premier aperçu quantitatif pour montrer l'ampleur et la variété du grec dans la Correspondance – et, par là-même, la pertinence de son (ré)examen. Pour le premier jeu de données exposé ici – à savoir les verbes et les participes – nous en établirons une typologie morpho-syntaxique en nous focalisant sur certaines formes en particulier, qui impliquent le processus de création verbale ou néologie. Le deuxième jeu de données – les citations grecques versifiées –, sera l'occasion d'interroger l'adéquation de ce concept polysémique de citation avec la constitution de notre corpus.

Références bibliographiques

Cicéron, Correspondance, lettres 1 à 954 (2021), introduction et notes de J.-N. Robert, commentaire de J.-N. Robert, traduit par L.-A. Constans, J. Bayet et J. Beaujeu, Paris, Les Belles Lettres

Adams J.N. (2003), Bilingualism and the Latin Language, Cambridge-New-York, Cambridge University Press

Aubert-Baillot S. (2021), Le grec et la philosophie dans la correspondance de Cicéron, Turnhout, Brepols

Compagnon A. (1979), La seconde main ou le travail de citation, Paris, Seuil

Dabène L. (1994), Repères sociolinguistiques pour l'enseignement des langues : les situations plurilingues, Vanves : Hachette FLE

Darbo-Peschanski C. (2004), « Les citations grecques et romaines », dans C. Darbo-Peschanski (dir.), La citation dans l'antiquité : actes du colloque du PARSA Lyon, ENS LSH, 6-8 novembre 2002, Grenoble, Million, p. 9-21

Firth J.R. (1957 [1951]), « Modes Meaning », dans Essays and Studies, The English Association, reprinted in Firth J.R. (1957), J. Papers in Linguistics 1934-1951, Oxford University Press, p. 190-215

Gumperz J.J. (1989b), Sociolinguistique interactionnelle, Paris, L'Harmattan

Habert B. (2000), « Des corpus représentatifs : de quoi, pour quoi, comment ? », dans Cahiers de l'Université de Perpignan 31, p. 11-58

Jackson J. (2014), « In utramque partem tum Graece tum Latine : Code-Switching and Cultured Identity in Cicero's Letters to Atticus », MA (non publié), University of Kansas

Nelson M. (2010), « Building a written corpus », dans O'Keeffe A. et Mc Carthy M. (dir.), The Routledge Handbook of Corpus Linguistics, Routledge, p. 53-65

Nicolas Chr. (2006), « Hôs ephat', dixerit quispiam », « comme disait l'autre » : mécanisme de la citation et de la mention dans les langues de l'Antiquité, Grenoble, ELLUG

Pruvost J. et Sablayrolles J.-F. (2019 [2003]), Les néologismes, Paris, PUF/Humensis, 4e éd.

Rochette B. (2012), « Problème du bilinguisme dans l'Antiquité gréco-romaine », dans Bertrand J., Boiley P., Genet J. et Pantel P. (dir.), Langue et histoire, Éditions de la Sorbonne

En ligne : https://books.openedition.org/psorbonne/83238?lang=fr

Steele R.B. (1900), « The Greek in Cicero's Epistle », dans American Journal of Philology 21, p. 387-410

Annexes

Tableau a – Étude de Steele (1900)

Quotations

Poets	Homère, Hésiode, Pindare, Eschyle, Sophocle, Aristophane et Euripide
Prose-writers	« The prose quotations are limited to **three authors** » : Platon, Thucydide et Épicure
Proverbs	« The larger part of them are to be found in the *Paroemiographi Graeci* (Leutsch and Schneidewin, 1839) » 2 catégories sont présentées : - *detached phrases* : « (...) which may be considered as colloquial expressions, though we do not know what their original associations may have been » (p. 403) « some have no verb expressed, and seem like **catch-phrases** » (p. 403) - *ciceronian phrases* : « the statement which may be considered **as Cicero's own contribution** to the Greek of Epistles are chiefly political, philosophical and geographical, with some entreaties and exclamations » (p. 403)

Individual words

Définition	« The citations in the Thesaurus of Stephanus have been taken as determining the occurrences of individual words, and they have been classified as occurring **only in Cicero, and first in Cicero** » *Adverbs* : « fifty-one different adverbs – forty-one **positive** forms, seven **comparative**, and four **superlative** » (p. 405) *Adjectives* : « one of the noticeable features in the use of adjectives is the number – fifty-four – **derived from verbs** » (p. 407) *Nouns* : 2 catégories (p. 408) - « found **only in Cicero**, though some of them have corresponding adjective or verbal form in Greek authors » - « seem to occur **first in Cicero** »
Catégories proposées	« Cicero uses 324 Greek nouns » ; le tableau ci-dessous (d'après Steele, 1900, p. 408) expose « the number of nouns with different endings, the most common prefixes, and the number of compounds of nouns and adjectives »

	ἀ-.	δυς-.	εὐ-.	φιλο-.	Prepositions.	Noun and Adj.	Noun and Noun.	Total Comp.	Total Number.
-ία,	16	4	11	3	17	10	8	69	84
-η,	—	–	—	–	15	—	2	17	38
-μα,	—	–	1	1	19	1	—	22	38
-σις,	1	–	—	–	31	—	—	32	48
-της,	—	–	1	–	1	1	2	5	9
Various,	3	–	—	–	26	5	4	38	107
	20	4	13	4	109	17	16	183	324

Verbs (p. 408-409) : 2 catégories également, similaires à celles des nom

Tableau a – Étude de Steele (1900)

Quotations

Poets	Homère, Hésiode, Pindare, Eschyle, Sophocle, Aristophane et Euripide
Prose-writers	« The prose quotations are limited to **three authors** » : Platon, Thucydide et Épicure

« The larger part of them are to be found in the *Paroemiographi Graeci* (Leutsch and Schneidewin, 1839) »

2 catégories sont présentées :

- *detached phrases* :

Proverbs

« (...) which may be considered as colloquial expressions, though we do not know what their original associations may have been » (p. 403)

« some have no verb expressed, and seem like **catch-phrases** » (p. 403)

- *ciceronian phrases* :

« the statement which may be considered **as Cicero's own contribution** to the Greek of Epistles are chiefly political, philosophical and geographical, with some entreaties and exclamations » (p. 403)

Individual words

Définition

« The citations in the Thesaurus of Stephanus have been taken as determining the occurrences of individual words, and they have been classified as occurring **only in Cicero, and first in Cicero** »

Adverbs : « fifty-one different adverbs – forty-one **positive** forms, seven **comparative**, and four **superlative** » (p. 405)

Adjectives : « one of the noticeable features in the use of adjectives is the number – fifty-four – **derived from verbs** » (p. 407)

Nouns : 2 catégories (p. 408)

- « found **only in Cicero**, though some of them have corresponding adjective or verbal form in Greek authors »
- « seem to occur **first in Cicero** »

Catégories proposées

« Cicero uses 324 Greek nouns » ; le tableau ci-dessous (d'après Steele, 1900, p. 408) expose « the number of nouns with different endings, the most common prefixes, and the number of compounds of nouns and adjectives »

	ἀ-.	δυς-.	εὐ-.	φιλο-.	Prepositions.	Noun and Adj.	Noun and Noun.	Total Comp.	Total Number.
-ία,	16	4	11	3	17	10	8	69	84
-η,	—	–	—	–	15	—	2	17	38
-μα,	—	–	1	1	19	1	—	22	38
-σις,	1	–	—	–	31	—	—	32	48
-της,	—	–	1	–	1	1	2	5	9
Various,	3	–	—	–	26	5	4	38	107
	20	4	13	4	109	17	16	183	324

Verbs (p. 408-409) : 2 catégories également, similaires à celles des noms

Tableau b – Étude de Jackson (2014)

Verbs, deliberation and identity (p. 10)

Dans le corpus des lettres à Atticus :

Catégories proposées

- « of the approximatively 760 times that Cicero code switches, 527 or 70% are single-word switches, which are **the most common type of code-switching** »
- « only 60 (11%) of those single-words switches are **verbs** »

Adjectives (p. 15)

« at 141 instances, there are more than twice as many adjectives as verbs in the category of single-word code-switches in the letters to Atticus »

Observations sur **le genre grammatical** (« the distribution of adjective gender »)

Nouns (p. 18)

« at 251 instances, nouns make up the largest portion of Cicero's single word code-switches »

Observations sur **le cas** (« their case distribution ») et sur **le genre grammatical**, par rapport aux adjectifs

Est également détaillé le cas des **noms propres** : « another 34, or 14%, of Cicero's single Greek nouns are proper nouns » (p. 23)

Quotations (p. 25)

Dans le corpus des lettres à Atticus :

- « there are 80 Greek quotations, of which 57 are **full quotations** – thas is, complete sentences or lines of poetry »
- « the other 23 are **partial**, cutting off in the middle of a thought and left for the reader to complete »

Cicero's longer Greek phrases and sentences (p. 35)

Jackson remarque : « even the multi-words Greeks phrases Cicero uses are usually just two or three words, but these phrases often provide additional insight into **how Cicero's Greek and Latin work together** »

Cette catégorie se manifeste en particulier :

- dans le cadre d'une démonstration philosophique (ex. *Att.* 9.4.1-2.)
- dans la conclusion de la lettre, « to express greetings between his son and Atticus' son » (ex. *Att.* 2.12.4.)
- lorsque Cicéron se trouve loin de Rome et qu'il demande à Atticus des renseignements sur ses affaires privées et publiques (ex. *Att.* 4.13.2.)

Tableau c – Étude de Adams (2003)

Catégories proposées	Description et exemples
Critical terms	« (…) **the most common type** of switching in the letters (at least those Atticus) consists of brief characterisations (usually by a single Greek adverb or sometimes by a short phrase) of someone's words »
Code-switching as form of coding or exclusion	« There was a concern about **the security of letters in antiquity**, in the absence of formal postal system, and here Cicero expresses his that letters may betray him » Exemples : *Att.* 2.20.3. ἀλληγορίαις ; *Att.* 7.7.1. ἐν αἰνιγμοῖς « He means partly but not exclusively the use of Greek substitute terms for **proper names** » Exemple : « (…) the frequent use of βοῶπις for *Clodia* »
Code-switching as distancing or euphemism	« A switch into a second language which is not the native language of the addressee may have **a distancing effect**. Thus switches into Greek are sometimes made for euphemistic purposes » Exemples : « (…) Cicero switches into Greek in alluding to characteristics of his own **which might be construed as faults** » « (…) **criticisms of others** (or references to their possible defects) are softened by switches into Greek »
Code-switching and proverbial or fixed expressions	« (…) **clichés** from one language may be known ever to monolingual speakers of another, and thus if worked into speech by a bilingual even when addressing a monolingual thay may serve to establish the solidarity of joint understanding » « (…) proverbs, particularly in the form of **literary tags**, in many cases belonged to **Greek literary culture**, and their use had much the same function as the use of technical terms from the various *artes* or of phrases from Greek literature : they placed the user and the recipient within **the hellenophile cultural *élite*** »
Code-switching and the « mot juste »	« The Roman were well aware of the existence of Greek words **difficult to render into Latin** » « Sometimes Greek offered **not merely** a single word appropriate to Roman concerns, but a whole group of cognates terms, nominal, verbal, adjectival and adverbial, which allowed a topic to be discussed in a more economical and varied way than might have been possible in Latin » Exemple : the derivatives of πόλις
Code-switching and medical terminology	« Whereas some disciplines, most notably rhetoric and philosophy, were being actively Latinised in the Republican period, medicine preserved **its Greek identity longer** » « The use of medical Greek in the letters evokes **the Greekness of the *ars***, though that said Cicero would hardly have code-switched under such circumstances in a public speech ; the topic or domain is **only** influential within the framework of the hellenising mixed language in which Cicero and Atticus communicated in private »
Special cases : the evocativeness of code-switching	« (…) the use of the proper Greek term by Cicero to evoke **an object or the like** » Exemples : « (…) using the Greek form of **the scultptor's name** » ; « Cicero often gives **the titles** of Greek literary works in Greek »

Automatiser l'extraction et le classement de séquences candidates à la catégorie des prépositions complexes en français

Ghayoung Kahng[1], Olivier Kraif[2], Denis Vigier[1]
[1] Laboratoire ICAR UMR 5191, Université Lyon 2
[2] Univ. Grenoble Alpes, LIDILEM, F-38000 Grenoble, France
[3] Laboratoire ICAR UMR 5191, Université Lyon 2

La présente contribution propose de nouvelles avancées en vue de relever l'un des défis que Stosic & Fagard (2019 : 8) jugent prioritaire pour toute étude sur les prépositions complexes, et cela en dépit des nombreux travaux déjà conduits sur la question : « en dresser la liste ».

Sous réserve qu'elle s'accompagne d'une réflexion critique sur les frontières de la catégorie et qu'elle soit conçue à grande échelle, nous pensons nous aussi qu'une telle liste revêtirait un intérêt majeur non seulement pour le traitement automatique de la langue (on disposerait d'outils à même d'affiner la segmentation des unités pour le traitement automatique du français contemporain) mais aussi pour les linguistes travaillant sur les prépositions complexes du français. Grâce à elle en effet, ces derniers disposeraient d'une sorte « d'arrêt sur image » leur permettant d'appréhender plus finement la dynamique interne propre à la catégorie des prépositions complexes. Le moteur de cette dynamique, on le sait, résulte d'une somme de processus de figement et de grammaticalisation qui conduisent ces dernières à s'échelonner suivant un graduum complexe entre agencements de mots très figés comme *à l'insu de, à cause de ….* (voire fusionnés et réanalysés en unités simples par les locuteurs contemporains : *depuis, parmi, …*) et agencements plus ou moins libres. On se trouve là face à ce que Melis (2003 : 115) nomme « l'imbrication du discours et du code et, plus en général, du caractère dynamique du système linguistique ».

Avant de présenter la méthode que nous nous proposons de suivre et qui mobilisera les outils de la linguistique quantitative outillée *via* le Lexicoscope (http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0), il convient de s'arrêter rapidement sur les travaux portant sur les prépositions en français afin de contextualiser notre propos et nos objectifs.

Depuis les années 90 tout particulièrement, de nombreux travaux ont été conduits sur les prépositions simples du français, que cela soit en synchronie ou en diachronie. Les prépositions complexes et les locutions prépositionnelles ont en revanche moins retenu l'attention de la communauté des chercheurs, même si l'on dispose à leur sujet de travaux notables auxquels nous ferons référence (en part. Gross 1981, 2006 ; Borillo 2001 ; Adler 2001 ; Fagard & De Mulder 2007 ; Stosic & Fagard 2019 ; Fagard, Pinto de Lima & Stosic (eds) (2020); Vigier & Kahng, 2022). Concernant cette sous-catégorie de prépositions, la terminologie varie. La démarche la plus fréquente consiste à distinguer les prépositions simples (*à, de, en, dans, contre …*) *versus* construites (*parmi, malgré, …*) d'une part, les prépositions complexes (*quant à, à cause de, à l'instar de, …*) *versus* les locutions prépositives d'autre part. Ces dernières sont formées de groupes de mots non entièrement

figés et vérifiant un nombre plus ou moins réduit de propriétés sémantiques et syntaxiques propres aux séquences libres : compositionnalité du sens, opérations d'insertion, de commutation, de transformation, …

La distinction entre prépositions complexes et locutions prépositives ne doit néanmoins pas masquer le fait que l'ensemble des séquences réunies dans ces deux catégories peuvent être rangées selon un continuum d'expressions plus ou moins figées rendant problématique sa discrétisation, opération pourtant nécessaire pour former deux sous-classes. De fait, la plupart des travaux jusqu'ici consacrés aux prépositions complexes et aux locutions prépositives se sont heurtés à cette question de la frontière entre les deux, comme le rappellent Stosic & Fagard (2019 : 13-14). Voilà pourquoi ces auteurs proposent de modifier la perspective traditionnellement adoptée par l'analyse et plaident pour une approche scalaire des catégories qu'ils traitent par le modèle de la théorie du prototype dans sa version standard (Kleiber 1990). Ainsi sont-ils conduits à proposer une grille multicritères énonçant 21 propriétés morphologiques, sémantiques, syntaxiques et fréquentielles attribuables aux prépositions complexes (désormais PrepComp) et permettant de ranger chacune des séquences examinée le long d'un continuum de plus ou moins grande prototypicalité vis-à-vis du noyau de la catégorie. Ce noyau est lui-même formé des séquences qui vérifient le plus grand nombre des propriétés identifiées. Vigier & Kahng (2022), adoptant l'approche prototypique proposée par Stosic & Fagard (2019), ont prolongé leur étude en proposant une grille multicritères alternative, plus réduite (15 tests au lieu de 21) et au pouvoir discriminant plus élevé.

Ces grilles permettent désormais de conduire des études linguistiques à même de classer les agencements de mots selon un continuum de séquences plus ou moins proches du noyau de la catégorie des PrepComp. Cependant, le processus mis en jeu, qui s'appuie sur de nombreux tests effectués manuellement, est long et ne permet pas de constituer rapidement de vastes listes. Or, comme le font observer Stosic & Fagard (2019 : 8), le nombre estimé des PrepComp est élevé : « Tandis qu'on compte généralement quelques douzaines de prépositions simples dans les langues qui ont cette catégorie, les estimations sont bien plus élevées pour les prépositions complexes – plusieurs centaines au moins (voir par exemple Borillo 1991, 1997 et Le Pesant 2006 pour le français, Huddelston & Pullum 2002 pour l'anglais). »

Nous nous sommes donc proposés d'automatiser le processus de création de listes de séquences candidates à la catégorie des PrepComp en nous appuyant fortement sur les travaux conduits en linguistique antérieurement et cités pour partie *supra*. Un des grands intérêts en effet de la proposition de Stosic & Fagard (2019) a été d'introduire parmi les propriétés caractéristiques des PrepComp la mesure, en corpus, de la force des liaisons internes entre les mots constitutifs de la locution. De tels calculs recourent à des indices largement utilisés aujourd'hui dans les Sciences Humaines et Sociales (t-score, z-score, Log-Likelihood, spécificités de Lafon etc., Seretan 2011) et permettent d'évaluer la significativité statistique de la mesure d'association entre un pivot et tel ou tel de ses co-occurrents sélectionnés dans une fenêtre de cooccurrence. Vigier & Kahng (2022), emboîtant le pas à Stosic & Fagard (2019), ont proposé de systématiser la perspective ouverte par ces derniers en proposant une approche plus précise, plus fine et plus systématique des mesures effectuées. Néanmoins, leur approche peut-être considérée comme semi-automatique en ce que les calculs conduits, quoique mobilisant un environnement informatique (diverses fonctionnalités statistiques implémentées sur plateforme SketchEngine et appliquées à la base de donnée FrTenTen17), sont effectués manuellement ainsi que le report des valeurs obtenues.

La démarche présentée dans cette communication franchit un pas supplémentaire en proposant une méthode entièrement automatisée. Le corpus sur lequel nous travaillons est constitué de la réunion de plusieurs sous-corpus : des articles de presse issus du *Monde* pour la période 1980-2019, des articles du média en ligne *Reporterre* qui publie depuis 2008 et enfin des articles scientifiques en SHS publiés entre 1990 et 2020, pour un total d'environ 280 millions de tokens. Ce vaste corpus est intégré dans l'outil « Lexicoscope 2.0 » qui constitue notre plateforme de travail.

Dans un premier temps nous présenterons les nouvelles fonctionnalités que nous avons implémentées dans cet outil afin notamment d'automatiser les propositions formulées par Vigier & Kahng (2022) pour la mesure systématique de la force des liaisons internes entre les mots constitutifs de la séquence testée. Ces fonctionnalités s'appuient sur un calcul des mesures d'association interne entre chaque élément périphérique et le noyau de la séquence. Par exemple, pour la séquence "à l'instar de", analysée dans le Lexicoscope comme un sous-arbre gouverné par *instar* et comportant 3 éléments régis, les prépositions *à* et *de*, ainsi que le déterminant *le*, nous calculons successivement les mesures d'association entre chacun de ces 3 éléments et le reste de l'expression. Nous émettons l'hypothèse que plus une séquence est figée, plus elle aura tendance à « verrouiller » ses éléments périphériques. A contrario, dans une combinaison libre, on suppose que les éléments périphériques sont les premiers à être sujet à variation. La moyenne de ces mesures permet ainsi de situer chaque séquence candidate à la catégorie des PrepComp sur un continuum d'expressions plus ou moins cohésives.

Dans une deuxième partie, nous étudierons comment les mesures d'association proposées dans la littérature (t-score, information mutuelle spécifique, *loglike*, etc.) sont corrélées aux observations effectuées à l'aide des tests linguistiques. Nous nous appuierons sur un corpus limité de séquences dont la proximité au noyau prototypique des PrepComp a été mesurée manuellement par l'application des tests linguistiques d'une partie des grilles de Stosic & Fagard (2019) et de Vigier & Kahng (2022). En particulier, nous chercherons à limiter l'effet de certains biais, qui font que les mesures d'association ont tendance à favoriser, pour certaines, des unités de haute fréquence, et pour d'autres les combinaisons plus rares (Evert, 2007).

Dans la dernière partie, nous exposerons sur quels principes nous avons isolé un vaste corpus de séquences candidates à la catégorie des PrepComp en recourant en particulier à la liste des patterns de cette catégorie dont on dispose désormais en français (voir les typologies de L. Melis (2003 : 107-108) et de Stosic & Fagard (2019 : 15-16) notamment). Puis nous présenterons les listes que nous avons obtenues pour chacun de ces patterns et nous commenterons la valeur de ces résultats.

Références

Adler, S. (2001), « Les locutions prépositives : questions de méthodologie et de définition », Travaux de linguistique, 42-43,1, 157-170.

Borillo, A. (1991): « Le lexique de l'espace : prépositions et locutions prépositionnelles de lieu en français », in: Tasmowski, L. & A. Zribi-Hertz (éds.): Hommage à N. Ruwet. Communication & Cognition, Gand, pp. 176–190.

Borillo, A. (1997), « Aide à l'identification des prépositions composées de temps et de lieu », Faits de langue, 9, 173-184.

Borillo, A. (2001), « Il y a prépositions et prépositions », Travaux de linguistique, 42-43 (1-2), p. 141-155.

Evert, S. (2007). « Corpora and collocations ». In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. Berlin : Mouton de Gruyter.

Fagard, B. & W. De Mulder, (2007) « La formation des prépositions complexes : grammaticalisation ou lexicalisation ? » *Langue française*, 156, 9–29.

Fagard, B., J. Pinto de Lima & D. Stosic (eds) (2020): *Complex adpositions in European Languages*, De Gruyter Mouton.

Gross, G. (1981), « Les prépositions composées », in C. Schwarze (éd.), *Analyse des prépositions*, 3e colloque franco-allemand de linguistique théorique, p. 29-39.

Kleiber, G. (1990), La sémantique du prototype. Catégories et sens lexical, PUF : Paris.

Le Pesant, D. (2006), « Classification à partir des propriétés syntaxiques ». *Modèles linguistiques*, 53, pp. 51–74.

Melis, L. (2003), *La préposition en français*, Paris : Ophrys.

Seretan, V. (2011), *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Dordrecht : Springer.

Stosic, D. & Fagard, B. (2019), « Les prépositions complexes en français. Pour une méthode d'identification multicritère », *Revue romane*, 54 (1), p. 8-38.

Vigier, D., Kahng, G. (2022), « Catégoriser les prépositions complexes en français », 8ème *Congrès Mondial de Linguistique Française*, Orléans, France

Russian Learner Corpus and Spelling Issues

Irina KOR CHAHINE[1], Ekaterina UETOVA[2]
[1] Laboratoire BCL, Université Côte d'Azur, CNRS, Nice, France
[2] Technological University, Dublin, Ireland
Irina.kor-chahine@univ-cotedazur.fr, euetova@gmail.com

Introduction

Databases which collect students' productions and are commonly called learner corpora, can be used for pedagogical purposes. In this domain, learner corpora[67],—whether they are of a native L1 or a foreign language L2—are intended to enhance students' literacy and grammatical skills (like CROW[68], for instance). In pedagogical uses of learner corpora, spelling skills are normally not subject to particular attention. However, spelling is a part of linguistic command, and studies on native L1 spelling occupy an important place in pedagogical research (see among others Lederlé 2011, Estienne 2014 for L1 French).

The lack of attention to spelling issues in L2 learner corpora studies has to be viewed from the vantage point of second language acquisition (SLA). Despite the large bibliography devoted to the acquisition of a foreign language, few studies focus on the acquisition of spelling (see among others, Ibrahim 1978, Tesdell 1982, Luelsdorff 1986, 1991; Cook 1997, 2001; Morris 2001, Howard *et al*. 2012, Khansir 2013, Brosh 2015, Llombart-Huesca 2018). However, beyond its semiotic aspect, spelling itself represents a valuable material in understanding acquisitional processes of foreign languages. And in this respect, learner corpora open new perspectives for spelling studies.

Based on Russian L2 learner corpora, our study relies on the analysis of spelling error frequencies, among both non-Russian-speaking subjects and heritage learners[69]. It seeks to trace trends by conducting a cross-sectional analysis of these errors. It also classifies spelling errors according to mechanisms (transposition, substitution, insertion, and omission) that come into play.

Corpus and methodology

Corpus

In this talk, we propose to focus on spelling errors in a written corpus made up of productions L2 by Russian language learners. We are interested in the handwritten errors produced by foreign learners of Russian. Our corpus consists of written productions by French students

[67] A useful table gathering some learner corpora is available at the website of the Centre for English Corpus Linguistics: Learner Corpora around the World. Louvain-la-Neuve: Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (last accessed 01.08.2022).

[68] https://crow.corporaproject.org/

[69] The term of *heritage learners* refers to speakers who use two languages at the same time, with one being reserved for the family environment and the other being used in a linguistic environment outside the family (work, study, social life). About heritage speakers see also Ortega 2020.

learning Russian, collected and annotated in the French subcorpus of the Russian Learner Corpus (http://www.web-corpora.net/RLC/). The French subcorpus of texts mainly consists of written student works (42 083 words; from A1 to C1 foreign language levels, as well as from native speakers) mainly comes from the University Cote d'Azur (Nice), the University Lyon 2 and, in Paris, Sorbonne University (Paris IV) and the École Normale Supérieure (ENS, rue d'Ulm). Details concerning the corpus are reported in Table 1.

Language background	Language level	Ratio	Number of words	Number of texts	Average number of words per text	Standard deviation in number of words per text
Foreign Learners	A1	5.74%	2 416	22	109.82	87.16
	A2	27.74%	11 673	103	113.33	77.55
	B1	16.24%	6 836	45	151.91	65.19
	B2	6.00%	2 527	16	157.94	54.45
	C1	4.13%	1 740	9	193.33	87.92
FLs Total		59.86%	25 192	195	129.19	78.53
Heritage Learners	B1	0.44%	187	3	62.33	31.48
	B2	2.30%	966	6	161	58.21
	C1	12.84%	5 402	34	158.88	78.75
HLs Total		15.58%	6 555	43	152.44	77.81
Native speakers		24.56%	10 336	43	240.37	166.24
TOTAL		100.00%	**42 083**	281	149.76	104.81

table 1. : Annotated corpus (token counts) in the corpus according to French students' level and group

Two groups took part in the study: non-Russian-speaking subjects (foreign learners: FLs) at the A1-C1 level, according to the CEFR[70] proficiency level, learning Russian as an L3 and often as a fourth language (L4) or fifth language (L5), and French-Russian learners (heritage learners: HLs) at the B1-C1 level. To these two groups is added a group of Russian students who attend French universities through international exchanges and who formed a control group composed of native speakers (NSs).

Methodology

Our study is based on the analysis of 1816 spelling errors[71] of Russian learners distributed by levels and student groups. The spelling errors come from the multi-lingual L1 corpus[72]. The obtained quantitative data clearly show that the rate of spelling errors decreases and, therefore, the difficulties posed by this question gradually disappear and are proportional to the language level.

Chart 1[73] shows that the frequency of spelling errors among foreign students (FLs) at the A1 level is quite moderate at the beginning of learning, and that these errors decrease considerably from the B1 intermediate level, the "starting" level for any HLs proficient in

[70] Common European Framework of Reference for Languages, see www.coe.int.

[71] A sample of working tables is available at https://russianwheel.univ-cotedazur.fr/research.html.

[72] The multi-L1 corpus means that students have various L1s but have spent several years in French environment. For more detail on students' profile see Kor Chahine & Uetova 2021.

[73] Spelling level ratios on the y-axis are a ratio of the errors to the total number of words within each proficiency level. For example, there are 2416 words and 187 errors on A1 level of foreign learners subcorpus: 187 / 2416 = 0,08.

Russian as a heritage language. At the B1 level, the error ratio of these heritage learners (HLs) largely exceeds the number of errors of foreign subjects. This fact may at first seem quite remarkable. However, the very high number of errors made by HLs at the beginning of learning Russian (B1) can be explained by the fact that HLs mainly have a mastery of oral Russian that has been acquired in the family environment and the fact that their writing skills are usually very low.

Chart 1 also shows a remarkable progression in spelling mastery, as this sample of HLs significantly reduces the gap in the number of errors made at the next language level (B2). As a result, there is almost the same rate of errors at the end of learning (C1 level) for both HLs and FLs, even if the latter are still very competitive (their error rate is lower) in spelling since FLs are more aware of the spelling question from the beginning of learning.

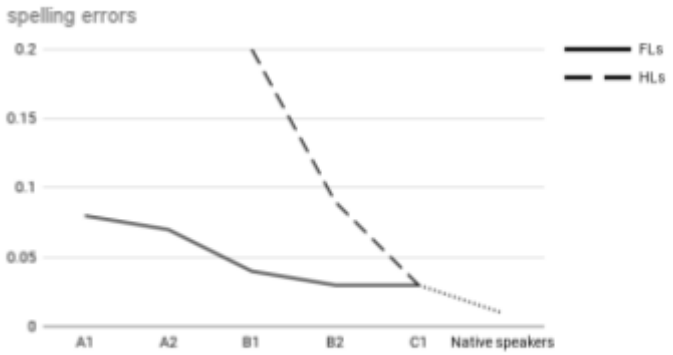


Chart 1. Spelling Errors by Level and Group: FLs and HLs Compared to Native Speakers

While most spelling errors involve standard words of Russian, some errors – like errors in hyphenation, errors in word or non-word spacing, and errors in capitalization – represent a deviation from spelling standards as well. Having a very low frequency rate (0.29%), these errors have little impact on the semantic integrity of a word, demonstrating a lack of knowledge with respect to purely orthographic rules (in hyphenation or capitalization, for instance) or a miscomprehension of words' boundaries (word spacing). We will not detail them in this study.

Results

Through the four mechanisms that operate in the errors of learners of Russian in a French-speaking environment (transposition or inversion, insertion, omission and substitution), it appears that the identified errors were motivated either by the close linguistic context or by various noncontextual factors. Among these factors it is worth mentioning the importance of long-term memory (in particular, word memory and visual memory of writing at both the interlinguistic and intralinguistic levels) and the interference of other languages, both the L1 mother tongue and languages already learned or being acquired.

Error decoding is a multifactorial process; thus, each error is committed under the influence of a parameter bundle. However, the recurrence of some errors leads us to think about their systemic nature. Although this study focused on francophone learners, recent investigations involving other corpora (Spanish, English, German, Finnish, Kazakh) have shown that most of

the found errors were characteristic of all learners of Russian, regardless of their origin (Ogneva 2018). Nevertheless, further research is needed to confirm the universality of the spelling errors addressed in this study.

The results of our quantitative and qualitative analysis of spelling errors of FLs and HLs represent interesting findings for research not only in acquisition and linguistics, but also in instructional design methods. The results of this study could be used for the design of specific purposes.

Bibliography

de Angelis, G. (2007). *Third or Additional Language Acquisition*, series "Second Language Acquisition" 24. Clevedon, Buffalo, Toronto: Multilingual Matters Ldt, 152 p.

Alonso Alonso, R. (Ed.). (2016). *Crosslinguistic influence in Second Language Acquisition*. Bristol: Multilingual Matters, 230 p.

Benson, M. (1973). "The spelling of past passive participles in Russian", *The Slavic and East European Journal*, 17 (4), 433-436.

Benson, M. (1993). "A note on Russian orthography", *The Slavic and East European Journal*, 37 (4), 530-532.

Brosh, H. (2015). "Arabic Spelling: Errors, Perceptions, and Strategies", *Foreign Language Annals*, 48 (4), 584-603.

Cook, V. J. (1997). "L2 Users and English Spelling", *Journal of Multilingual and Multicultural Development*, 18 (6), 474-488.

Cook, V. J. (2001). "Knowledge of writing", *International Review of Applied Linguistics in Language Teaching*, 39 (1), 1-18.

Češko, L. A. (ed.). (1956/1963). *Russian Orthography*, Moscow: Učpedgiz, 1956; translated by T. J. Binyon, edited by C. V. James, 1963, Oxford/London/New York/Paris: Pergamon Press.

Elson, M. J. (1975). "Morphological aspects of Russian Spelling", *The Slavic and East European Journal*, 19 (1), 85-90.

Estienne, Fr. (2014). *Dysorthographie et dysgraphie: 300 exercices: comprendre, évaluer, remédier, s'entraîner*, Collection "Orthophonie". Issy-les-Moulineaux: Elsevier Masson, 182 p.

Hamers, J. F., Blanc, M. H. A. (1989). *Bilinguality and Bilingualism*. Cambridge: Cambridge University Press.

Howard, E. R. (2012). "Can yu rid guat ay rot? A Developmental Investigation of Cross-Linguistic Spelling Errors Among Spanish-English Bilingual Students", *Bilingual Research Journal*, 35, 164-178.

Hristova, D.S. (2011). "*Velikij i mogučij olbanskij jazyk:* The Russian Internet and the Russian Language", *Russian Language Journal*, 61, 143-162.

Ibrahim, M. H. (1978). "Patterns in Spelling Errors", *English language teaching journal*, 32 (2), 207-202.

Khansir, A. A. (2013). "Error Analysis and Second Language Writing", *Theory and Practice in Language Studies*, 3 (2), 263-270.

Kor Chahine, I., Uetova, E. (2021). "From Error Annotation to Quantitative Analysis: Patterns in Russian Language Learning", *Russian Language Journal*, 2021, 71 (3), 39-70. ⟨hal-03376956⟩

Lederlé, E. (Ed.) (2011). *Les troubles du langage écrit: regards croisés*. Ortho Edition, 430 p.

Ljaševskaja, O., Šarov, S. (2009). *Častotnyj slovar' sovremennogo russkoja jazyka (na materialax Nacional'nogo russkogo jazyka)*. <Dictionnary of frequency of Modern Russian (based on National Corpus of Russian)>. Moskva: Azbukovnik. Open access: http://dict.ruslang.ru/freq.php?

Llombart-Huesca, A. (2018). "Understanding the Spelling Errors of Spanish Heritage Language Learners", *Hispania*, 101 (2), 211-223.

Luelsdorff, P. (1986). *Constraints on error variables in grammar bilingual misspelling orthographies*. Amsterdam: J.Benjamins Pub. Co.

Luelsdorff, P. (1991). *Developmental orthography*. Amsterdam/Philadelphia: J.Benjamins Pub. Co.

Morris, L. (2001). "Going Through a Bad Spell: What the Spelling Errors of Young ESL Learners Reveal about Their Grammatical Knowledge", *The Canadian Modern Language Review*, 58 (2), 274-286.

Ogneva, A. (2018). "Spelling errors in L2 Russian: evidence from Spanish-speaking students", *Estudios interlingüísticos 6*, 116-131.

Ortega, L. (2020). "The study of heritage language development from a bilingualism and social justice perspective", *Language Learning, 70*(S1), 15-53.

Robinson, P., Ellis, N. C. (Eds.) (2008). *Handbook of cognitive linguistics and second language acquisition*. New-York & London: Routledge, 566 p.

Rothman, J., Cabrelli J. (2009). What variables condition syntactic transfer? A look at the L3 initial state. *Second Language Research*, *25*(4), 1–30.

Simonchyk, A., Darcy, I. (2018). "The effect of orthography on the lexical encoding of palatalized consonants in L2 Russian", *Language and Speech*, 61 (4), 522-546.

Tesdell, L.S. (1982). *ESL spelling errors: a taxonomy*. Iowa State University. https://lib.dr.iastate.edu/rtd/7903

De nouvelles écritures pour documenter la part langagière de milieux didactiques : le cas des ateliers de la voie professionnelle en Guyane

Patricia Lambert[1], Sophie Alby[3], Zeynab Badreddine[5], Victor Corona[2], Ingrid de Saint-Georges[4], Anna Ghimenton[7], Justine Lascar[2], Abdelhak Qribi[6], Anna Claudia Ticca[2]

[1] UMR 5191 ICAR, ENS de Lyon
[2] UMR 5191 ICAR, CNRS
[3] UMR 8202 SEDYL, Université de Guyane
[4] MLING, université du Luxembourg (DHUM)
[5] Advanced Video Based Research (AVBRE)
[6] MINEA, Université de Guyane
[7] UMR 5596 DDL, Université Lyon 2

patricia.lambert@ens-lyon.fr, sophie.alby@univ-guyane.fr, Zeynab.Badreddine@AVBRE.com, victor.corona@ens-lyon.fr, Ingrid.DeSaintGeorges@uni.lu, Anna.Ghimenton@univ-lyon2.fr, justine.lascar@ens-lyon.fr, abdelhak.qribi@univ-guyane.fr, anna.ticca@ens-lyon.fr

Introduction

Cette présentation fera état de l'avancement de l'élaboration d'un corpus complexe et de son organisation en vue de la production de discours numériques et plurisémiotiques dans le cadre de la diffusion des résultats de la recherche.

Par « complexe », nous entendons ici : i) la pluralité des modes de production de données induite par une approche ethnographique (journal de terrain, entretiens, observations directes, recueil de documents, etc.) ; ii) l'accomplissement du travail en équipe pluridisciplinaire (sciences du langage, sciences sociales, sciences de l'éducation) ; iii) la diversité des types de matériaux empiriques à sélectionner, organiser, traiter, analyser.

Cette contribution s'inscrit dans le cadre de LaBør (*Language at the Borders of work and school*, Labex Aslan 2022-2024), un projet qui porte sur le langage dans la formation et la socialisation des lycéens et lycéennes de la voie professionnelle en Guyane, collectivité territoriale française située en Amérique du Sud. Ce contexte où les deux tiers des élèves ont une autre langue de première socialisation que la français (Léglise, 2017) et dans lequel la poursuite des études après la 3ème vers la voie professionnelle est « priorisée par les familles » (projet académique, 2018-2021) représente un site à forts enjeux pour l'étude des formes langagières dans les activités de formation et de professionnalisation. Partant, l'objectif premier de cette recherche, initiée récemment, est de documenter la part langagière de la formation professionnelle au sein des « ateliers » des lycées professionnels, via l'élaboration d'un répertoire d'études de cas (Passeron & Revel, 2005).

Afin de documenter ce type de milieu didactique encore rarement étudié, différentes spécialités sont ciblées (hôtellerie-restauration, métiers du bois, de la petite enfance,

gestion-administration, spécialités agricoles). L'organisation du corpus ainsi que le traitement et l'analyse des données s'appuieront sur les fonctionnalités du logiciel Transana® Multiuser, outil multimodal et collaboratif (Badreddine & Woods, 2014). Le choix de cet outil est notamment fondé sur l'objectif de production, en fin de programme, d'un support numérique qui rendra compte des résultats sous une forme multimodale (photos, vidéos, textes audio et scripturaux). Nous souhaiterions que cette ressource puisse être non seulement utile pour la recherche mais aussi pour la formation d'acteurs de la formation professionnelle (enseignants, formateurs, personnels d'encadrement).

Ce sont donc les enjeux et défis posés par tout le processus de ce nouveau type d'écriture (scientifique et de médiation de savoirs) que nous souhaitons mettre en discussion à un stade relativement précoce de la recherche.

Corpus et méthodologie

Aux frontières des univers de l'école et du travail, les « ateliers » occupent une place essentielle dans l'expérience et dans la formation professionnelle des lycéens et lycéennes de la voie professionnelle. Ils sont pourtant encore très peu étudiés par la recherche et assez méconnus de manière plus générale. Pousser la porte d'ateliers de différentes familles de métiers, faire porter notre attention sur les savoirs et les pratiques saisis dans leur dimension langagière, c'est pour nous une voie pertinente pour connaître et comprendre ce qui se vit, ce qui se travaille et ce qui s'apprend au sein de ces espaces éducatifs. Nos objectifs sont les suivants :

- Élaborer des connaissances sur les ateliers des lycées professionnels en Guyane
- Documenter la part langagière de la formation professionnelle
- Constituer un répertoire d'études de cas, utile pour la recherche et pour la formation
- Contribuer au partage de connaissances sur la formation professionnelle en atelier

Pluridisciplinaire, ce projet n'en est pas moins porteur d'enjeux théoriques et méthodologiques pour les sciences du langage. Il représente en effet une contribution au développement d'une linguistique sociale de la formation professionnelle (Filliettaz & Lambert, 2019) qui impose certaines exigences en matière de méthodologie de recueil et d'analyse des données langagières. Pour atteindre nos objectifs, le dispositif de recherche se fonde ainsi sur une enquête de terrain qui comprend les étapes et opérations suivantes :

- Organisation de l'enquête à l'aide d'outils de collaboration avancés (Huma-Num)
- Enquête de terrain collective et, autant que possible, collaborative : mise en œuvre de différents modes de production de données dont l'observation de Travaux Pratiques ; identification interdisciplinaire de situations-clés ; enregistrements audio/vidéo et transcriptions
- Création d'une base de données multisémiotique référencées via le logiciel d'analyse Transansa® MU
- Analyse multidimensionnelle et construction analytique des cas
- Restitution aux acteurs du terrain sous la forme d'écritures numériques

Le tableau ci-dessous offre un aperçu (maximaliste) des spécialités et diplômes ciblés et des focus analytiques qui pourront être opérés.

Formations	Types de données et points de focus	
Métiers du bois (CAP ou Bac pro)	- Corpus des référentiels de la formation	Approche contrastive (Est/Ouest) sur les

Métiers du bois (CAP ou Bac pro)	- Corpus entretiens élèves suivis (2 par classe) et enseignants ateliers - Corpus d'observations en atelier : Travaux pratiques (TP) observés (partiellement filmés) pendant deux semaines consécutives (16h x 2) **> Focus : mono/ plurilinguisme, multimodalité, résolution de panne ou problème**	plans linguistique et pédagogique
CAP ATMFC (Assistant technique en milieux familial et collectif)	- Corpus des référentiels de la formation - Corpus entretiens élèves suivis (2 à 4) et enseignant - Corpus d'observations en atelier : Travaux pratiques (TP) observés (partiellement filmés) pendant deux semaines consécutives (16h x 2) **> Focus : jeux de rôle, multimodalité, langue française, discours réflexifs sur la pratique**	Approche contrastive des référentiels : - par spécialité - entre spécialités - entre niveaux de diplômes
Bac pro Gestion-Administration	- Corpus des référentiels de la formation - Corpus entretiens élèves suivis (2 à 4) et enseignant - Corpus d'observations en atelier : Travaux pratiques (TP) observés (partiellement filmés) pendant deux semaines consécutives (16h x 2) **> Focus : langue française/pratiques plurilingues, littératie**	
Baccalauréat CGEA (Professionnel Conduite et Gestion de l'Exploitation Agricole – horticole et animale)	- Corpus des référentiels de la formation - Corpus entretiens élèves suivis (2 à 4) et enseignants ateliers - Corpus d'observations en atelier : Travaux pratiques (TP) observés (partiellement filmés) pendant deux semaines consécutives (16h x 2) **> Focus : littératie-numératie, multimodalité**	

Tableau 1. Dispositif d'enquête 2022-2024

Résultats

Sur les deux missions annuelles prévues en 2023 et 2024, la première aura lieu entre mars et mai de cette année. Nous ne sommes donc pas encore en mesure de faire état du corpus et des résultats. Nous serons néanmoins à même de présenter une première organisation de la base de données au moment des journées d'étude et de soumettre à la discussion un certain nombre de pistes et de questions qui pourront porter sur différents aspects du travail.

L'élaboration de cette base de données impose déjà des discussions scientifiques extrêmement fécondes autour des critères utilisés pour la constitution de collections de situations-clés, de classes de phénomènes à analyser et de leur définition. Nous souhaitons que ce travail permette à terme, dans la lignée d'autres projets, la mise à disposition pour la communauté de jeux de données réexploitables pour la recherche mais aussi dans le cadre de formations.

Au-delà des questions scientifiques et techniques qui se posent habituellement pour la constitution d'un corpus aussi complexe (création d'une base de données multi-sémiotique référencée), l'équipe se rend donc particulièrement attentive à l'anticipation de modes de restitution sous différents formats et notamment en explorant de nouvelles écritures scientifiques. Nous envisageons par exemple la réalisation de capsules vidéo augmentées et celle d'environnements 360 avec des hotspots sur le logiciel 3D vista liant les différents modalités et formats des éléments du corpus permettant des expériences de type immersif. Le collectif qui mène l'enquête étant interdisciplinaire, nous serons également amenés à questionner ce qui peut différencier la construction de notre corpus d'autres démarches de constitution de base de données vidéographiques, élaborées dans des perspectives moins hétérogènes.

Notre contribution fera donc état des premiers avancements dans le processus de constitution de la base de données complexes ; elle se propose également de présenter une illustration des diverses modalités d'écritures scientifiques possibles en fonction des données recueillies lors

de la mission en printemps 2023 et des potentialités des terrains investigués. Elle reviendra enfin sur les plus-values épistémologiques liées à l'élaboration d'un corpus dans une perspective interdisciplinaire.

Références bibliographiques

Académie de la Guyane (2018). *Projet académique 2018-2021*.

Badreddine, Z., Woods, D. (2014). L'usage de Transana pour l'étude de l'action conjointe et de la co-construction du sens en classe de sciences. *Recherche en didactiques*, 1.17, p. 93-111.

Filliettaz L., Lambert P. (2019). La formation professionnelle, un point aveugle de la linguistique sociale ? *Langage et société*, 168, p. 15-47.

Léglise, I. (2017). Les langues parlées en Guyane : une extraordinaire diversité, un casse-tête pour les institutions. *Langues et cité*, DGLF - Observatoire des pratiques linguistiques, 2017, Les langues de Guyane, pp.2-5.

Passeron, J.-C., Revel, J. (eds). (2005). *Penser par cas*. Paris : Éditions de l'École des hautes études en sciences sociales.

A corpus-based syllabus of Italian collocations

Francesca La Russa[1], Maria Roccaforte[2] et Veronica D'Alesio[3]
[123] Sapienza Università di Roma
francesca.larussa@uniroma1.it, maria.roccaforte@uniroma1.it, veronica d'alesio@uniroma1.it

Introduction

Lexical combinations are central to language learning because they can be processed quickly (Siyanova-Chanturia, 2015) and their use gives the idea of fluency in production (Nattinger & DeCarrico, 1992). However, the acquisition of L2 phraseological competence is often difficult for learners. This is particularly true for collocations, "sequences of words which tend to occur in stable and privileged combinations" (Simone, 1990: 440). The semantic transparency of collocations facilitates their understanding and makes them difficult to notice. Since collocations are often not highlighted in language courses, learning them is even more difficult because students do not notice and assimilate them as complex lexemes (Bini *et al.*, 2007). As a matter of fact, in Italian L2 syllabuses, vocabulary is often presented as a list of single words and the phraseological dimension is usually absent.
To fill this gap, our work aims to design a corpus-based syllabus of Italian collocations that indicates the collocations that should be taught/learned at different proficiency levels.

Corpus and methodology

Corpus

Following the model of the *English Vocabulary Profile*[74], collocations were extracted from a learner corpus which provides reliable data on learners' authentic use of the language.
The CELI corpus (Spina *et al*., 2022) was chosen. It collects 3041 written texts produced by learners of Italian L2 who passed the CELI exams[75] (levels B1, B2, C1, C2). The main corpus is made up of four different sub-corpora. Each sub-corpus collects the written productions corresponding to a given level. Further information on the composition of the corpus can be seen in table 1[76].

Level	N. texts	Token	Media token	Type	TTR	Sentences	Token x sentence
B1	1212	156.612	129,21	7.397	18,69	13.514	11,58
B2	840	152.251	181,25	9.519	24,39	8.438	18,04
C1	585	149.859	256,16	12.546	32,4	7.508	19,95
C2	404	149.892	371,01	14.153	36,55	7.196	20,82
Total	**3041**	**608.614**				**36.656**	

Table 1. Composition of the CELI corpus

[74] https://www.englishprofile.org/wordlists
[75]CELI (Certificati di Lingua Italiana) are certificates of Italian language competence released by Perugia Foreigners' University.
[76] Adapted and translated from Spina *et al*., (2022: 127).

Methodology

Since a given collocation is often used at different proficiency levels, the following criteria were adopted to assign it to the most suitable level:

- frequency of the collocation in the *Perugia Corpus* (PEC) (Spina, 2014) - which collects written and oral texts produced by native speakers;
- number of occurrences of the collocation in the CELI subcorpora;
- presence of the collocates in the lexical lists of the *Profilo della lingua italiana* (Spinelli & Parizzi, 2010);
- topic.

Table 2 synthesize the procedure that was adopted to assign each collocation to a given proficiency level.

Coefficient of usage in native speakers' production	Based on their coefficient of usage (a measure that combines frequency and dispersion through the different textual genres in the corpus, cf. Juilland & Chang Rodriguez, 1964) in the PEC corpus, a distinction was made between collocations belonging to high, medium, or low frequency range. ● collocations in the high frequency range should be assigned to level B1 or level B2; ● collocations in the medium frequency range should be assigned to level B2 or level C1; ● collocations in the low frequency range should be assigned to level C1 or C2.
Number of occurrences in the CELI subcorpora	Between the two proficiency levels indicated by the frequency range, the collocation was assigned to the level in which it occurs more often. If the number of occurrences in the two levels is the same or similar or the collocation is already consistently used at the lower level, it was assigned to the lower level.
Italian Profile	When criteria 1 and 2 give contrasting information, the Italian Profile lexical lists were checked, and the collocation was assigned to the level to which the words that make up the collocation belong.
Topic	It was double checked if the collocations assigned to a given proficiency level address topics that are relevant to that level.

Table 2. Procedure adopted to assign collocations to a proficiency level

The following examples may help clarifying the procedure:

- *trovare lavoro*, 'find a job': is highly frequent. Therefore, it should be assigned to level B1 or B2. It is used 58 times at level B1 and 39 times at level B2, thus it was assigned to level B1.
- *avere diritto*, 'have right to': belongs to the high frequency range Therefore, it should be assigned to level B1 or B2. However, it is never used at level B1, it is used 5 times at level B2, 40 times at level C1 and 20 times at level C2. The word *diritto* does not appear in the A1 to B2 lexical lists of the Italian Profile, we can therefore assume that it is learned at an advanced level. The topic addressed is that of socio-political structures, thus it is more relevant for C level learners. Since the collocation is more often used at level C1, it was assigned to this level.

Results

The result is a syllabus in which 946 Italian collocations are organized according to the proficiency level they should be taught/learned and the topic they refer to.

References

Bini, M., Pernas, A., Pernas, P. (2007). Apprendimento e insegnamento collocazioni dell'italiano. Con i NUNC più facile, in *Corpora e linguistica in rete*, M. Barbera, E. Corino & C. Onesti (eds.), 323–333, Perugia: Guerra Edizioni.

Juilland, A., Chang-Rodriguez, E. (1964). *Frequency Dictionary of Spanish Words*. The Hague, Mouton & Co.

Nattinger, J., DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford University Press.

Simone, R. (1990). *Fondamenti di linguistica*. Bari: Laterza.

Siyanova-Chanturia, A. (2015). On the 'holistic'nature of formulaic language, *Corpus Linguistics and Linguistic Theory*, 11(2): 285-301.

Spina, S. (2014). Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione, in *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014. vol. 1*, R. Basili, A. Lenci, B. Magnini (eds.), 354-359. Pisa: Pisa University Press.

Spina, S., Fioravanti, I., Forti, L., Santucci, V., Scerra, A., Zanda, F. (2022). Il corpus CELI: una nuova risorsa per studiare l'acquisizione dell'italiano L2, I*taliano LinguaDue*, 14 (1):116-138.

A corpus-based study on mistakes in English prepositions made by French and Russian learners

Lebedeva Iuliia
ICTT, Université d'Avignon
iuliia.kasianova@alumni.univ-avignon.fr

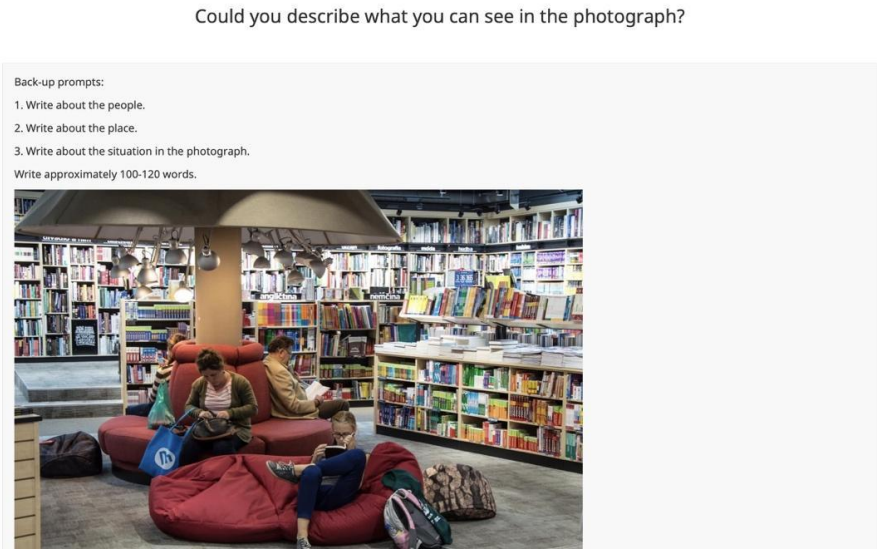
Introduction

The study aims to examine the language skills of French and Russian learners of English, with a special focus on English prepositions. The participants of the experiment are French and Russian students who learn English for special purposes. The participants have Beginner – Advanced levels of English. The average age of the respondents is 17-24 years old. The total number of the participants is 176.

Corpus and methods

The corpus is compiled from French and Russian texts written by the participants. The data elicitation task is a survey on the online platform "Lime Survey". The survey consists of two types of questions. The first part includes questions which give some general information on the participants' backgrounds, age, language level, etc. This part aims to recover metadata, with a view to categorising responses.

At the second stage of the survey French and Russian learners have to describe a photo in English (pic.1). This format of the task is widely used not only in Cambridge exams which Russian students take to prove that they have achieved B1 or B2 in English, but also in the unified state exam which Russian students have to take in order to enter the university. The time of the task is limited. Students have to spend 10 minutes writing texts. All the respondents are asked not to use any dictionaries or resources while doing the task.



Picture. 1 Writing task on the platform “limesurvey”

Results

First, mistakes in prepositions are analyzed at different levels of English. The table below shows the number of participants by level of English.

Level of English	**Number of participants**
Advanced (C1)	8
Beginner (A0)	9
Elementary (A1)	40
Pre-intermediate (A2)	54
Intermediate (B1)	33
Upper-intermediate (B2)	20
Unknown	11

table.1 The number of participants by level of English

The chart below shows the average number of mistakes made at each level of English in two groups of the participants.

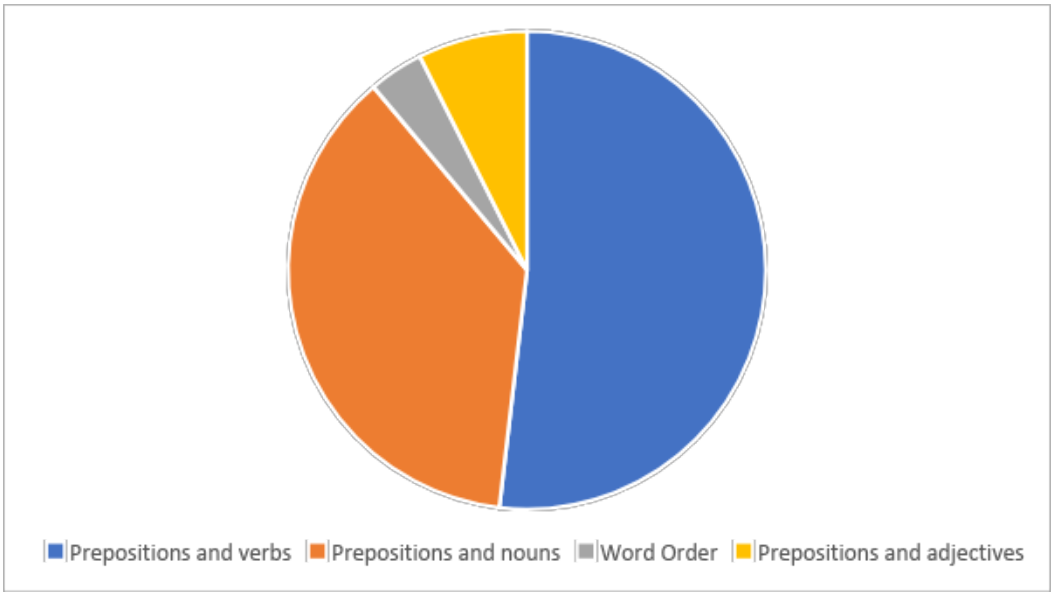


Chart. 1 Mistakes in English prepositions made by French and Russian learners at different levels of English

The chart above represents the number of mistakes in the groups of participants. The blue bars represent French respondents, whereas the orange bars show the number of mistakes made by Russian learners. The absence of mistakes is connected either to the fact that the participants did not make any mistakes or the fact that there are not any participants who have that particular level of English. The number of mistakes is not distributed equally between French and Russian learners from A1-C1. French learners tend to make more mistakes at pre-intermediate and intermediate levels, whereas Russian students have difficulty in using prepositions when they have levels A1 or B1. As we see both groups made almost the same number of mistakes at level B1.

At the next stage of our research, we summarize the data obtained in the survey and classify mistakes in the writing task. According to the analysis French and Russian respondents experience difficulty in using English prepositions in spatial contexts, e.g.:

French student: *On this picture we can see, four peoples. At the first plan a young girl who is reading a book.*

Russian student: *In the foreground of the photo is a young girl, most likely she went after school to read her favorite book in the library, since next to her lies a school bag she sits on an ottoman in jeans and a T-shirt.*

French student: *At the second plan a woman who is search on her bag next to her, and a old man who is raeding a book.*

Russian student: *The place on the photo is a library, we can see a lot of books.*

French student: *on this picture, ican see four people, in a bibliotecary, i can see one teenager, an old man who can have sixty old years*

Russian student: *Man is reading and the girl on the background sitting on her telephone.*

The most numerous group of mistakes is found in the examples where students have to indicate the location of the objects in the photo, e.g., the participants made mistakes in the phrases “in the background” and “in the foreground” when they were describing the part of

the photo which is the farthest from the viewer. The correct phrases were substituted with such expressions as “at the background, on the backside, on the background, on front side, on the foreground, in the front of this picture etc.”

Also, we divide the category of mistakes in spatial contexts into several syntactic types of mistakes, depending on the part of speech which prepositions are collocated with, e.g.:

- prepositions and nouns,

French student: *In the picture i can look 4 person who read books they are in library with more book*

Russian student: *The people on photograph art sittig on sofa and some kind of armchair.*

table 1. : prepositions and adjectives

French student: *The readers must be here because they were looking for calm to read, the place has bright lights to allows a comfortable reading time.*

Russian student: *They are in the library, which is full off books or in a booking shop, in which it is possible to read. This place is very suit.*prepositions and verbs

French student: *The woman with a blue bag search something in her other bag*

Russian student: *Look in the picture. L can see people who sit in the library. They are busy with their own affairs.*

table 2. : the wrong word order

French student: *On this picture we can see four peoples, two off them are reading a book, one on is phone and the last one is a lady who searches in his bag something*

Russian student: *Behind her sits a man and two women*

The categories of mistakes are analyzed in the two groups of respondents. The charts below provide information on the number of mistakes made by French and Russian participants in each category.

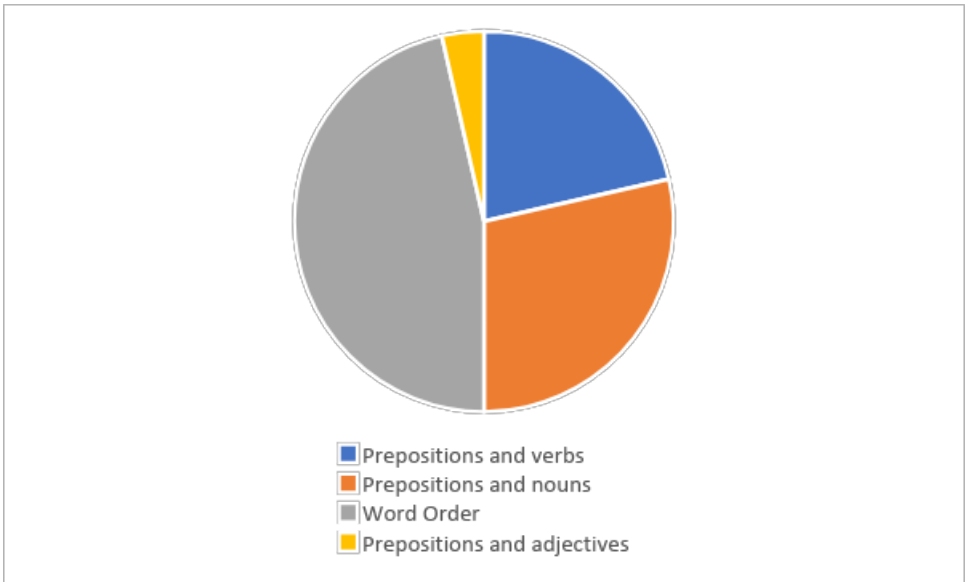


Chart.2 The number of mistakes made in English prepositions by French learners

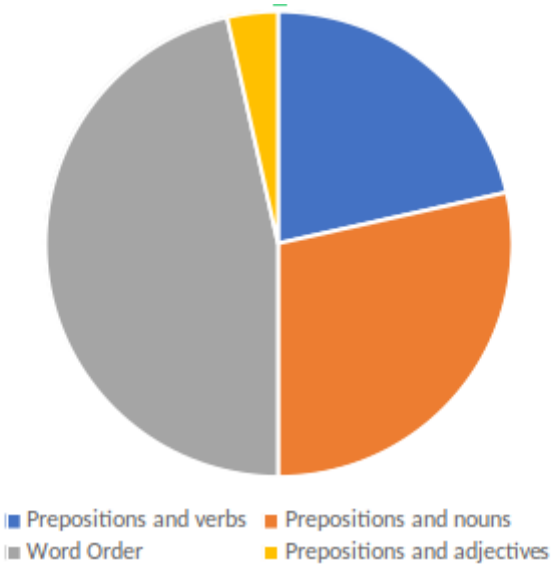


Chart. 3 The number of mistakes made in English prepositions by Russian

The pie charts show that Russian respondents tend to make more mistakes in the category of the wrong order, whereas French learners misuse prepositions in collocations with verbs.

Conclusion

We suppose that there are different reasons why French and Russian learners make mistakes in prepositions. There are external factors causing mistakes in English prepositions apart from the participants' mother tongue. The factors will be analyzed in further research.

References

Andrea Tyler and Vyvyan Evans: The Semantics of English Prepositions. Spatial Scenes, Embodied Meaning and Cognition (2003). Cambridge: Cambridge University Press, 254.

Arleo A. R.R. Jordan, English for academic purposes: A guide and resource book for teachers (1998). *Cahiers de l'APLIUT*. 17(4):78.

Dale R, Anisimoff I, Narroway G. HOO (2012): A Report on the Preposition and Determiner Error Correction Shared Task. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. Association for Computational Linguistics; 54-62.

Gries STh, Stefanowitsch A, eds. (2006) Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis. In: Trends in Linguistics. Studies and Monographs [TiLSM]. Mouton de Gruyter; 1-17.

Han, N., Tetreault, J., hwa Lee, S., and young Ha, J. (2010) Using an error-annotated learner corpus to develop and esl/efl error correction system. In *Prcoeedings of LREC 2010*. Valletta, Malta.

Jordan RR. English for Academic Purposes: A Guide and Resource Book for Teachers (2005). 1st edition. Cambridge University Press.

Tummers J, Heylen K, Geeraerts D. Usage-based approaches in Cognitive Linguistics: A technical state of the art. 2005;1(2):225-261.

Zavarykina, L.V. (2018) Teaching English for Specific Purposes in Russian Universities: A Case of Moscow School of Social and Economic Sciences. *Vysshee obrazovanie v Rossii Higher Education in Russia*. Vol. 27. No. 11: 62-70.

Constitution et exploration du corpus de discours scientifique oral en français pour une étude phraséologique

Chaeyoung Lee
Laboratoire LIDILEM, Université Grenoble Alpes
chaeyoung.lee@univ-grenoble-alpes.fr

Introduction

La pratique de la langue dans les discours scientifiques suscite un intérêt croissant au sein des Sciences du Langage tant pour sa description et son analyse linguistiques que pour son application à la didactique de la langue sur objectifs académiques dans le milieu universitaire (p. ex. *English for Academic Purposes* (EAP) et *Français sur Objectifs Universitaires* (FOU)). Parmi les différents phénomènes linguistiques spécifiques au genre scientifique, on peut noter notamment la présence d'un ensemble de formules phraséologiques partagées communément par les locuteurs de la même « communauté scientifique » (Swales, 1990). Cette phraséologie caractéristique du discours scientifique est étudiée principalement dans deux objectifs : mettre en évidence les particularités phraséologiques propres au genre scientifique, qui le distinguent de la langue générale, et développer des méthodes d'aide à la maîtrise des phraséologismes fréquents chez les étudiants dans leur production académique.

En français, la plupart des recherches menées sur ce sujet jusqu'à présent sont principalement concentrées sur les registres écrits, tels que les articles de recherche, les mémoires de Master ou les thèses de Doctorat et d'HDR (Tutin, 2014 ; Tran, 2014 ; Yan, 2017, Jacques & Tutin, 2018). Cependant, les communications orales lors d'évènements scientifiques, tels que les conférences, les journées d'étude et les séminaires, jouent également un rôle essentiel dans la construction et la diffusion de nouvelles connaissances scientifiques (Jacques, 2017). C'est dans ce contexte, plus précisément en constatant le manque de ressources exploitables pour les registres scientifiques oraux, que le projet EIIDA (2012-2017) a élaboré un corpus multi-lingue (anglais, français et espagnol) de discours scientifiques écrits (articles de recherche) et oraux (présentations de conférence) (Carter-Thomas & Jacques, 2017). L'un des objectifs de ce projet est d'étudier l'impact du mode de communication (écrit ou oral) sur l'utilisation de la langue dans le genre scientifique, car l'environnement discursif spécifique à l'oral – tel que la présence directe de l'allocutaire, la simultanéité et l'immédiateté des interactions entre les participants du discours, l'utilisation de supports visuels (diapositives ou exempliers) ou les contraintes du temps (Jacques, 2017) – peut exercer une influence déterminante sur la manière dont le chercheur s'exprime pour présenter ses activités de recherche et communiquer ses connaissances acquises.

Relevant du domaine d'étude sur la *phraséologie transdisciplinaire* (Tutin, 2014), notre étude s'intéresse plus particulièrement sur les expressions phraséologiques utilisées pour prendre en compte l'allocutaire dans la construction du discours scientifique oral. À cet effet, nous avons conçu un nouvel objet phraséologique, « formules discursives », qui englobe différents types d'unités phraséologiques de nature hétérogène en termes de figement : des courtes *phrases*

préfabriquées des interactions (Tutin, 2019) de type continu et fixe (p. ex. *on va dire*, *vous voyez*, *si vous voulez*, etc.) aux *routines sémantico-rhétoriques* (Tutin & Kraif, 2016) de type assez flexible, qui sont des patrons lexico-syntaxiques permettant diverses variations. Dans le cadre de ce projet de thèse, notre objectif est de développer, à travers les formules discursives, l'une des premières représentations du paysage phraséologique du genre des communications scientifiques orales en français. Cette initiative pourrait nous conduire, à l'avenir, à appliquer notre typologie fonctionnelle des formules, ainsi que notre corpus, à la didactique du FOU, en particulier dans le but d'élaborer des méthodes de formation pour aider les étudiants à mieux se préparer à une présentation scientifique orale.

Corpus et méthodologie

Corpus

Ayant pour objectif d'étudier la phraséologie du discours scientifique oral, nous avons créé un corpus de communications scientifiques orales, que nous appelons le CComSciO (Lee, 2022a, 2022b), comprenant trois disciplines des Sciences Humaines et Sociales : la linguistique (dont une partie est issue du corpus EIIDA), les sciences de l'information et de la communication et la didactique des langues. La composition détaillée de ce corpus est présentée dans le tableau ci-dessous.

	Sous-disciplines	Nombre de communications	Durée	Nombre de mots
1	Linguistique	20	8h 51m 22s	97,229
2	Sciences de l'information et de la communication	20	8h 12m 45s	92,273
3	Didactique des langues	20	9h 01m 24s	103,827
Total		**60**	**26h 05m 31s**	**293,329**

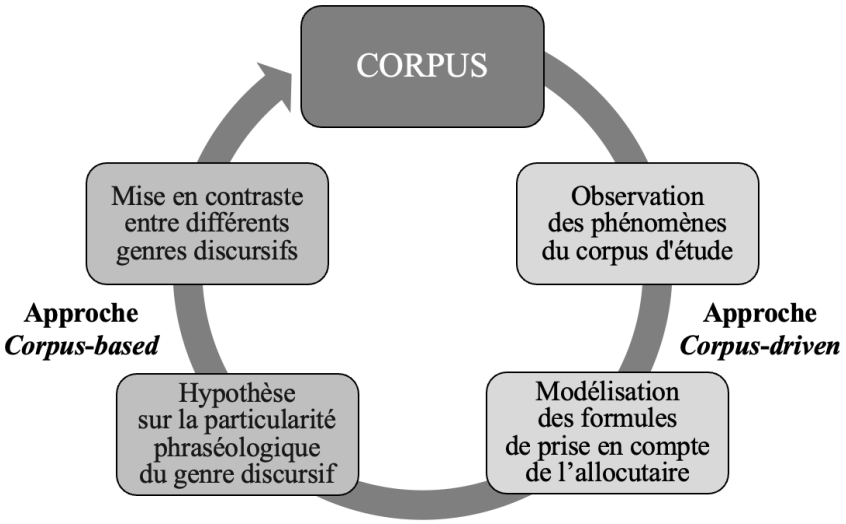
table 2. : **Description de la composition de notre corpus d'étude**

Le corpus est constitué de 60 présentations scientifiques orales, soit empruntées de la partie française orale du corpus EIIDA, soit enregistrées en audio sur place, soit téléchargées à partir de sites d'archives de podcasts scientifiques en ligne. Nous avons choisi les communications en tenant compte de plusieurs critères tels que la durée, la période de production, le nombre de locuteurs, etc., pour assurer la pertinence et la cohérence des données constituant le corpus. Après avoir obtenu l'accord des locuteurs concernant la collecte et l'utilisation des données, les communications ont été transcrites en orthographe standard, segmentées en constructions verbales selon la convention du corpus Gold de l'ORFEO, et alignées entre les données son et texte sur le logiciel ELAN.[77] La vérification du travail de la transcription et de la segmentation par des chercheurs natifs a eu lieu à plusieurs reprises.

[77] Nous tenons à remercier l'équipe du projet EIIDA d'avoir partagé leur sous-corpus français oral, le consortium CORLI d'avoir financé la construction de notre corpus, et nos collègues d'avoir contribué à la transcription, la segmentation, et l'alignement son-texte automatique et manuel.

Méthodologie

Cette étude adopte principalement une double approche : l'une de *corpus-driven* et l'autre de *corpus-based*. La démarche de notre étude est schématisée dans la figure suivante.



Approche double de linguistique outillée dans notre étude

En un premier temps, nous adoptons l'approche *corpus-driven*, en commençant par observer dans le corpus d'étude, des phénomènes phraséologiques récurrents, ainsi qu'en essayant de comprendre le fonctionnement discursif général des communications scientifiques orales, un genre relativement peu étudié en français de nos jours. C'est ainsi que nous avons développé notre concept de *formules discursives* dans le but de l'appliquer à l'analyse phraséologique de notre étude. Cependant, nous n'excluons pas l'approche *corpus-based*, car nous envisageons d'effectuer des analyses comparatives entre le genre des communications scientifiques orales de notre corpus (CComSciO) et un autre genre, qu'il soit écrit ou oral. Cette démarche vise à valider notre hypothèse selon laquelle les communications scientifiques constituent un genre discursif distinct avec son propre paysage phraséologique.

Pour explorer les formules de prise en compte de l'allocutaire dans notre corpus d'étude, nous avons d'abord procédé à leur repérage manuel, en nous limitant aux formules construites sur un prédicat verbal. Ensuite, nous avons effectué une annotation fonctionnelle[78] des formules, en utilisant notre typologie métadiscursive, et nous avons identifié différentes caractéristiques syntaxiques, lexicales, sémantiques et/ou énonciatives de nos formules. Cela nous a permis de les modéliser par la suite, plus précisément afin de former des patrons complexes relevant du type flexible de notre définition des formules discursives.

[78] Nous avons procédé, sur un échantillon du corpus, à l'interannotation avec nos collègues natifs du français pour comparer les choix d'annotation individuelle, en valider un pertinent et affiner notre typologie fonctionnelle.

Cette étape d'annotation, d'identification et de modélisation est en cours sur des fichiers .xlsx, exportés depuis le logiciel ELAN, et ils seront réimportés dans le même logiciel une fois que les étiquettes d'annotation auront été définitivement validées.

Résultats et perspectives

Pour cette communication lors des 11ème Journées de Linguistique de Corpus, nous souhaitons tout d'abord présenter notre typologie des formules de prise en compte de l'allocutaire à l'oral scientifique, que nous avons divisée en trois catégories : 1) la catégorie **métalinguistique et métaénonciative** (la reformulation, la réparation, l'approximation métaénonciative, ...), 2) la catégorie de **structuration discursive** (l'annonce du plan de la présentation, les références à d'autres communications, le rappel, la limitation, ...), 3) la catégorie **interactionnelle** (la mise en valeur, le co-constat, l'appui sur des connaissances déjà partagées, ...). Inspirée du modèle de guidage du lecteur de Ji (2022) sur le discours scientifique écrit, notre typologie repose sur la notion de métadiscours, employée dans différents modèles expliquant le fonctionnement du discours académique (Hyland, 2005 ; Ädel, 2010). Nous fournirons aussi quelques exemples de formules surreprésentées dans chaque fonction, ainsi qu'un bref aperçu de la distribution des fonctions dans notre corpus d'étude.

Parallèlement à cette typologie, nous souhaitons également présenter les résultats principaux de nos deux analyses comparatives entre genres discursifs.

- Pour la comparaison entre les communications scientifiques orales (le CComSciO) et les articles de recherche (le TermITH), nous nous concentrons sur les formules à fonction de renvoi (inter-/intra-)discursif, aussi appelé la *navigation discursive* (Ji, 2022). L'objectif est de mettre en évidence la particularité du processus scientifique consistant à renvoyer à d'autres discours ou à des éléments du même discours à l'écrit et à l'oral. La première analyse révèle une présence significative de certaines fonctions spécifiques à l'oral telles que les *références à d'autres communications* (p. ex. *on en a parlé ce matin* ; *ça a été dit*) qui témoignent du dynamisme de l'environnement discursif dans une communication ou le *rappel négatif* (p. ex. *on vous en a pas parlé*), reflétant la simultanéité de la production du discours dans ce genre malgré la planification préalable.
- Pour la comparaison entre les communications scientifiques orales (le CComSciO) et les interactions en entretien semi-directif (le CFPP2000), dont une partie a déjà été réalisée par Lee (2022b), nous nous concentrons spécifiquement sur les formules construites sur le verbe *dire*. En effet, ces formules constituent l'un des phénomènes particuliers de la langue parlée. Par exemple, l'analyse de Lee (2022b) portant sur les formules les plus surreprésentées (*on va dire*, *je veux dire*, *je te/vous dis*, …) met en évidence deux résultats principaux : 1) la fonction de commenter l'approximation du locuteur dans le choix d'un terme ou d'une expression est beaucoup plus fréquente dans le discours scientifique ; 2) les formules d'insistance, *je te dis* ou *je vous dis*, sont complètement absentes dans les communications scientifiques orales où la rhétorique de la mise en valeur des éléments se réalise de manière plus formelle (p. ex. *j'insisterai sur le fait que ...*, *l'idée à retenir c'est que ...*).

En tant qu'une des premières analyses sur la phraséologie dans le discours scientifique, notre travail de thèse ouvre la voie à plusieurs pistes d'approfondissement. Premièrement, au-delà des formules discursives de type verbal et de la notion de prise en compte de l'allocutaire, il serait intéressant d'explorer le genre des communications scientifiques orales pour d'autres types de formules phraséologiques : par exemple, les marqueurs d'atténuation (*je pense*, *je crois*, *me semble-t-il*) ou les formules de type pseudo-clivé comme *ce qui est* ADJ *... c'est que*.

Ces examens permettraient de mieux appréhender la diversité des formules phraséologiques présentes dans ce registre scientifique oral.

Ensuite, notre typologie fonctionnelle des formules pourrait être appliquée dans le cadre de la didactique du français parlé académique. Par exemple, il est envisageable de développer des formations universitaires visant à améliorer les compétences en communication scientifique orale en français pour les doctorants étrangers. Cela contribuerait à approfondir les stratégies communicatives des étudiants, ainsi qu'à faciliter leur intégration dans le milieu universitaire francophone. Nous pourrions également envisager de mettre notre corpus d'étude, CComSciO, à disposition sur une plateforme numérique à des fins pédagogiques. Les usagers pourraient ainsi y effectuer des recherches sémantiques par fonction, qui leur permettent de consulter les exemples de formules de notre typologie et leur contexte d'utilisation. Cette ressource serait particulièrement utile pour les étudiants se préparant une présentation académique, car elle leur fournirait des références concrètes et authentiques concernant les phraséologismes oraux fréquemment utilisés dans un contexte académique.

Références bibliographiques

Ädel, A. (2010). Just to give you kind of a map of where are we going: A taxonomy of metadiscourse in spoken and written academic English. *Nordic Journal of English Studies* 9(2). 69-97.

Andersen, H. L. (2007). Marqueurs discursifs propositionnels, *Langue française* 154, 13-28.

Carter-Thomas, S. & Jacques, M.P. (2017), Interdisciplinary and interlinguistic perspectives on Academic Discourse: The mode variable. Introduction to the special issue on the French EIIDA project, *CHIMERA. Romance Corpora and Linguistic Studies* 4(1), 1-11.

Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*, New York, Continuum.

Jacques, M.P. (2017). La structuration textuelle en discours scientifique : comparaison oral/ écrit. *CHIMERA. Romance Corpora and Linguistic Studies* 4(1), 89-115.

Jacques, M.P. & Tutin, A. (éds) (2018). *Lexique transversal et formes discursives des sciences humaines*, Londres, ISTE Éditions.

Ji, Y. (2022). *Les routines de guidage du lecteur dans les écrits scientifiques en français*, Thèse de Doctorat, Grenoble Alpes Université.

Lee, C. (2022a). Quelques observations phraséologiques sur la co-construction des savoirs dans le discours scientifique oral en français, dans M. Hagafors, L. Heiden & L. Tarrade (éds), *ICODOC 2021 : le savoir au prisme du langage. Acquisition, transmission, manifestations*, *SHS Web of Conference* 146, 05003.

Lee, C. (2022b). Formules parenthétiques en *dire* et leur fonctionnement discursif dans les communications scientifiques orales, *Langue Française* 216, 13-28.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*, Cambridge, Cambridge University Press.

Tran, T.T.H. (2014). *Description de la phraséologie transdisciplinaire des écrits scientifiques et réflexions didactiques pour l'enseignement à des étudiants non-natifs : application aux marqueurs discursifs*, Thèse de Doctorat, Université de Grenoble Alpes.

Tutin, A. (2014). La phraséologie transdisciplinaire des écrits scientifiques : des collocations aux routines sémantico-rhétoriques, dans A. Tutin & F. Grossmann (éds), *L'écrit scientifique : du lexique au discours. Autour du Scientext*, Rennes, Presses Universitaires de Rennes, 27-44.

Tutin, A. (2019). Phrases préfabriquées des interactions : quelques observations sur le corpus CLAPI, *Cahiers de lexicologie* 114, 63-91.

Tutin, A. & Kraif, O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents, *Lidil* 53, 119-141.

Yan, R. (2017). *Étude des constructions verbales scientifiques dans une perspective didactique : utilisation des corpus dans le diagnostic des besoins langagiers en FLE à l'aide des techniques de TAL*, Thèse de Doctorat, Université Grenoble Alpes.

Exploration textométrique d'un corpus annoté et analyse discursive des évolutions du genre compte rendu de conseils de réunion en diachronie

Virginie Lethier[1], Frédérique Sitri[2], Emilie Née[2], Grigoriy Manucharian[3], Ilaine Wang[4]
[1] Laboratoire ELLIADD, Université de Franche-Comté
[2] Laboratoire CEDITEC, Université Paris-Est Créteil
[3] Laboratoire MoDyCo, Université Paris Nanterre
[4] Laboratoire ERTIM, Inalco
virginie.lethier@univ-fcomte.fr, frederique.sitri@u-pec.fr, emilie.nee@u-pec.fr, m.grigoriy@parisnanterre.fr, ilaine.wang@inalco.fr

Contexte de la recherche

Cette contribution s'inscrit dans le cadre général du projet ArchivU dont l'objectif est de saisir les transformations économiques, sociales et politiques qui affectent l'université depuis les années 1970 jusqu'à aujourd'hui. L'originalité de ce projet est d'envisager ces transformations sous un angle discursif, en se concentrant sur les discours que produisent les établissements dans des textes qui rendent compte au plus près de l'activité professionnelle multiforme qui s'y élabore et s'y déploie, telle qu'elle se configure dans les instances qui l'organisent, la régulent ou l'évaluent. En l'occurrence, sont interrogés deux genres de discours peu lus et souvent peu exploités, mais qui permettent justement d'appréhender la fabrique interne de l'institution universitaire : le compte rendu d'instance et le rapport de recherche de laboratoires (rapports d'activité et rapports d'auto-évaluation). Saisis dans leur dimension processuelle, en tant que pratiques professionnelles, ces genres donnent accès aux processus d'élaboration des normes de l'institution universitaire par ses discours, dans leur production, leur mode d'action ainsi que dans leur diffusion. Ces deux genres, qui représentent les deux grandes fonctions – formation et recherche – dévolues à l'université et permettent d'en saisir les évolutions, possèdent des points communs. Genres professionnels en tant qu'ils organisent et façonnent l'activité (Boutet 2014) ce sont aussi des genres administratifs qui, par leur mode de production, de diffusion et d'archivage, témoignent de l'idéal – imaginaire – de « transparence » qui anime les sociétés dites démocratiques. Produits à intervalles réguliers dans un cadre légalement contraint, ils postulent un allocutaire auquel on « rend compte » ou « fait rapport », c'est-à-dire qu'ils s'inscrivent dans un cadre contractuel vis-à-vis de cet allocutaire.

Notre choix de travailler à partir d'une perspective diachronique et longitudinale est à relier à une conception dynamique du genre du discours (Bakthine, 1984 ; Mellet et Sitri 2010). Revêtant une dimension heuristique, la notion de genre permet tout d'abord de constituer des corpus de textes associés par les locuteurs à une même fonction - que dénomme souvent le " nom de genre " - dans une pratique sociale. Mais nous concevons également le genre comme une catégorie énonciative : le choix d'un genre sélectionne des moyens langagiers et des modes d'énonciation qui " configurent " un sujet d'énonciation. Enfin, c'est à partir des

genres que l'on saisit des formes d'évidence que l'on peut rapporter à des déterminations interdiscursives (Dumoulin 2022, Sitri 2022, Dumoulin, Mellet, Sitri 2022).

Objectifs de la communication

Dans le cadre de cette contribution, nous nous proposons d'interroger les évolutions des comptes rendus de conseils (d'université/ d'administration) de l'université de Nanterre de 1984, date de la Loi Savary, à la loi ORE (2018).

Nous émettons l'hypothèse selon laquelle le genre du compte rendu configure un sujet du discours (Sitri, 2023) qui ne se réduit pas au "rédacteur" du texte et dont la figure évolue sous l'effet de différents facteurs extralinguistiques. Il nous semblerait légitime d'observer une mutation de ce sujet sous l'effet des réformes néolibérales de l'Enseignement Supérieur et de la Recherche, dont le mouvement s'accélère durant les années 2000. En effet, l'adoption, en 2007, de la loi relative aux libertés et responsabilités des universités (dite loi LRU) et le passage aux Responsabilités et Compétences Élargies (RCE) ont affecté en profondeur le fonctionnement et les missions de l'Université, ainsi que les rapports de force et les enjeux des conseils d'administration (voir entre autres Barats et al. 2018, Barthélemy et al. 2009 ou Harari-Kermadec 2021). Mettant au service de l'analyse du discours les méthodes des linguistiques de corpus et de la textométrie (Heiden et al., 2010), notre démarche d'analyse consiste à identifier de façon contrastive les variations d'observables formels entre les différents textes qui composent notre corpus, pour les interpréter en relation avec les évolutions historiques, sociales et politiques.

Traitant d'un corpus original s'inscrivant au carrefour du discours institutionnel et des écrits professionnels, notre contribution illustre ainsi les apports d'une exploration de données textuelles finement annotées pour l'analyse du discours.

Déroulé de la communication

Nous consacrerons le premier temps de notre communication à une présentation des caractéristiques du compte rendu et de notre corpus d'étude, ce dernier mettant en série 346 comptes rendus produits entre 1984 et 2018, pour environ 4.500.000 occurrences. Dans ce cadre, nous insisterons sur les spécificités énonciatives du genre du compte rendu (Sitri, à paraître) qui possède la propriété de « tenir lieu » de la réunion dont il est le compte rendu. Ce texte écrit suppose un locuteur **L** organisant la représentation des dires des locuteurs **l** tenus pendant la réunion. Le compte rendu de réunion est ainsi un genre « tout en RDA » (Authier-Revuz 2020 : 582-583), c'est-à-dire que le texte est interprété comme représentant la réunion en l'absence même de toute forme de discours rapporté.

Le deuxième temps de notre communication sera dédié à la présentation des principales tendances d'emploi des formes de RDA en diachronie. Ayant procédé à une exploration du niveau morphosyntaxique, nous avons en effet observé une évolution des pôles nominaux et verbaux, devant être interprétée comme l'indice d'une mutation des modalités de RDA. Le poids grandissant des verbes bénéficie en effet prioritairement aux verbes susceptibles d'opérer en contexte comme des introducteurs de discours autre. Il reflète ainsi le profil croissant du discours indirect. Nous rendrons compte de trois tendances fines de variations des verbes privilégiés pour mettre en scène le discours des participants au conseil : une tendance à l'explicitation de l'acte discursif à l'origine du contenu représenté ; un recul de la classe des verbes d'affect ; une raréfaction des modalisations des verbes introducteurs du

discours autre qui expriment l'attitude émotionnelle de **l**. Nous insisterons sur la façon dont le retour au texte nous a invité à émettre l'hypothèse selon laquelle l'identité de **l** serait source d'écarts à ces tendances, envisagées comme des normes endogènes au corpus.

Le troisième temps de notre communication sera ainsi consacré à rendre compte des analyses menées en vue de tester l'hypothèse selon laquelle la représentation du discours autre varie en fonction de l'identité de **l**, c'est-à-dire de son statut et de son genre. Après avoir très brièvement présenté l'annotation linguistique de haut niveau (Poudat, Landragin, 2017) qui a été déployée, nous montrerons la façon dont la mise en perspective de différentes mesures permet d'évaluer les déséquilibres quantitatifs de Représentation du Discours Autre en fonction du genre. Nous pointerons dans ce cadre que la sous-représentation des femmes au sein du conseil n'est, bien entendu, pas le seul facteur susceptible d'expliquer le déséquilibre observé entre la taille des séquences de RDA explicitement attribuées à une locutrice source et celle des séquences de RDA explicitement attribuées à un locuteur source.

Nous rendrons ensuite compte d'une analyse des spécificités de notre corpus partitionné par type de locuteurs en fonction de leur statut (exemple : représentants étudiants, personnels administratifs, présidence, etc.) et de leur genre (masculin/féminin). L'objectif de cette classe d'exploration est d'examiner si les différences de statut ou de genre peuvent être corrélées à la présence de certaines formes de RDA. On pourrait se demander, par exemple, si la fréquence des guillemets de modalisation autonymique n'est pas plus élevée dans des discours attribués à des locuteurs étudiants, guillemets encadrant des manières de dire que l'on peut considérer comme orales ou familières (1). On pourrait également se demander si le sémantisme du verbe introducteur de discours indirect ne peut être corrélé au statut de **l**, comme en (2) :

1. Mlle M. souhaite un acte politique condamnant la loi CESEDA, cette loi **« inepte et raciste ».** (2006-10-06, p. 36)
2. [Madame B.] **s'inquiète** par ailleurs du temps que les services informatiques vont pouvoir consacrer à l'avancement d'autres dossiers que celui de la mise en place de la carte IZLY.
3. Le Président **tient à rassurer** Madame B. en précisant que la feuille de route de la mise en place de la carte IZLY est bien établie et que tout devrait se passer sans difficultés. (2015-06-08)

Ces pistes, nécessairement exploratoires, seront soumises à la discussion.

Références bibliographiques

Authier-Revuz, J. (2020). *La représentation du discours autre. Principes pour une description,* De Gruyter.

Bakhtine, M. (1984, [1952-1953]). *Les genres du discours.* Dans *Esthétique de la création verbale*, Gallimard.

Barats, C., Bouchard, J., Haakenstad, A. (dir.) (2018). *Faire et dire l'évaluation. L'enseignement supérieur et la recherche conquis par la performance*, Presses des Mines.

Barthelemy, F., Beraud, A., Martin, M. (2009). La loi LRU a-t-elle modifié les distributions de pouvoir au sein des universités françaises ?. *Revue économique*, 6(60), 1469-1481. https://doi.org/10.3917/reco.606.1469.

Boutet, J. (2014). Les écrits au travail. Dans I. Laborde Milla, S. Plane, F. Rinck, F. Sitri (dir.), La formation aux écrits professionnels : des écrits en situation de travail aux dispositifs de formation, *Le discours et la langue,* 5(2)**,** 17-28.

Dumoulin, H. (2022). *Les théorisations du discours de Michel Pêcheux et Michel Foucault à la lumière du concept d'énonciation.* [Thèse de doctorat, Université Paris Nanterre].

Harari-Kermadec, H. (2019). *Le classement de Shanghai. L'université marchandisée*, Le bord de l'eau.

Heiden, S., Magué, J.-Ph., Pincemin, B. (2010). TXM : une plateforme logicielle open-source pour la textométrie – conception et développement. Dans S. Bolasco, I. Chiari et L. Giuliano (dir.), *10th International Conference on the Statistical Analysis of Textual Data –JADT 2010*, (p. 1021-1032). LED.

Mellet, C., Sitri, F. (2010). Nom de genre et institutionnalisation d'une pratique discursive : les cas du signalement d'enfant en danger et de l'interpellation parlementaire. Dans F. Neveu, V. Muni Toke et J. Durand (dir.), *2e Congrès mondial de linguistique française* (p. 781-795). EDP Sciences.

Sitri, F. Dumoulin, H., Facq-Mellet, C. (dir.) (2023). La fabrication discursive de l'université : comptes rendus et rapports scientifiques en diachronie. *Cahiers de praxématique*, 78. https://doi.org/10.4000/praxematique.8196

Sitri, F. (2023). La construction d'un « sujet du discours » dans les comptes rendus de CU/CA de l'université de Nanterre : une perspective énonciative ». *Cahiers de praxématique*, 78. https://doi.org/10.4000/praxematique.8196

Sitri, F. (2022). 'Genre de discours' et/ou 'formation discursive' : quelle articulation ?. Dans F. Neveu, S. Prévost, A. Steuckardt, G. Bergounioux et B. Hamma (dir.), *8e Congrès mondial de linguistique française*. EDP Sciences. https://doi.org/10.1051/shsconf/202213801001

Analyzing the Interdiscursivity in Microblog Marketing Discourse from the Perspective of Critical Genre Analysis: A Case Study of Uniqlo

Diqiao Li [1]
[1] School of Foreign Languages, South China University of Technology
202210187926@mail.scut.edu.cn

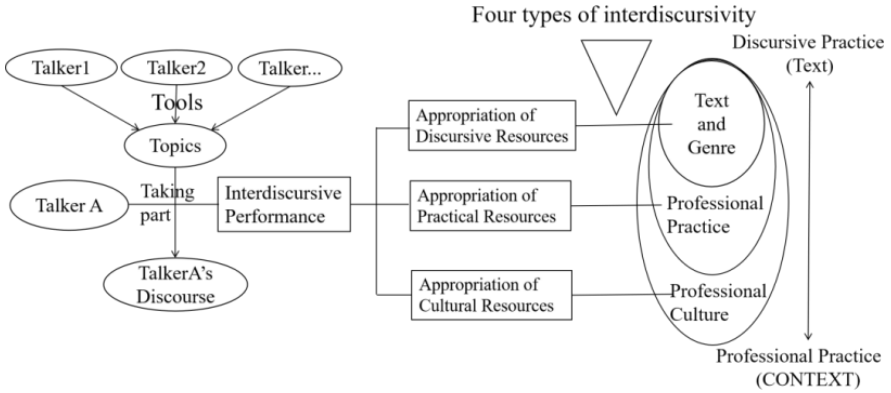
Introduction

Microblog, as a typical new media with a large number of users and huge page views, has attracted great scholarly attention to its increasingly diversified marketing activities and corresponding strategies. Taking Uniqlo's blogs on Weibo as the example, this paper, based on Bhatia's (2017) Critical Genre Analysis (CGA), Sernovitz's (2006) 5T word-of-mouth marketing model and Wu's (2012) classification of interdiscursivity, explores the interdiscursive performances of microblog marketing discourse from the perspective of text-genre, professional practice, and professional culture, so as to discover the business mechanism of pragmatic success in the new media marketing. It is found that microblog marketing discourse mainly consists of three types of genres: promoting products, building brands, and maintaining customer relations, among which other professional practices and cultures are interwoven through blending, embedding, switching, and chaining interdiscursivity, and gradually form three cultural tendencies of glocalization, popularization, and equalization. Such hybridization in these dimensions and flexible use of multimodal resources contribute to covering up the nature of marketing, and thus promoting the pragmatic success of marketing behaviors. This paper aims to provide an analytical lens for research on CGA and interdiscursivity, and help flesh out the social meaning of new media marketing.

Data Collection

Since this paper concerns the critical genre analysis of microblog marketing discourse, it randomly selects the blogs under the official account "优衣库_UNIQLO" on Weibo within a month from December 9, 2019 to January 9, 2020. Considering the highly interactive nature of Weibo, it includes not only the original texts, but also other blogs mentioned via words, images or hyperlinks, etc. After a manual inspection, 295 blogs with validity were finally extracted from its official account for further detailed analysis.

Analytical Framework



Analytical Framework for Interdiscursive in Microblog Marketing Discourse

Findings and Discussion

By analyzing the text-genre characteristics of Uniqlo's marketing discourse on Weibo, it is found that there are three salient genres, that is, promoting products, building brands, and maintaining customer relations.

No.	Move	Frequency	Occurrence	Position in the Blog	Optionality
1.	to take part in a topic	148	98.7%	Head	Compulsory
2.	to introduce a product	133	88.7%	Front part	Compulsory
3.	to offer the price	52	34.7%	Front part	Optional
4.	to bring forth a coupon hyperlink	21	14.0%	Latter part	Optional
5.	to present ways to purchase	134	89.3%	End	Compulsory
6.	to send an auspicious greeting	17	11.3%	End	Optional
7.	to invite interaction	36	24.0%	End	Optional

table 3. : Frequency and Proportion of Each Move for *Promoting Products*

No.	Move	Frequency	Occurrence	Position in the Blog	Optionality
1.	to publicize physical stores	15	18.8%	Front part	Optional
2.	to co-brand with other fields	52	65.0%	Front part	Optional
3.	to borrow celebrities' influence	13	16.2%	Front part	Optional
4.	to offer preferential treatment	61	76.2%	Latter part	Optional

table 4. : **Frequency and Proportion of Each Move for *Building Brands***

No.	Move	Frequency	Occurrence	Position in the Blog	Optionality
1.	to broadcast live-streaming videos	13	28.9%	Front part	Optional
2.	to launch "Koi Draws"	24	53.3%	Front part	Optional
3.	to invite to apply for the membership	8	17.7%	Front part	Optional

table 5. : **Frequency and Proportion of Each Move for *Maintaining Customer Relations***

Regular virtual observation of Uniqlo's marketing activities, together with preliminary analysis of data collected from Weibo, helped to discover that various types of linguistics manipulations and re-adoptions of semiotic as well as cultural resources is a distinct and salient feature of new media marketing.

To be specific, in the text-genre dimension, it is found that microblog marketing discourse is usually comprised of three kinds of genres. The genre of promoting products with the most moves manifests that how to introduce their products effectively constitutes the kernel for entrepreneurs to deal with; the genre of building brands with relatively fixed moves that the soft power construction is still relatively lagging; and the genre of maintaining customer relations that social platform provides much more resources to retain customers in a highly interactive way. For example, in addition to initiating a topic themselves, entrepreneurs could reshape the marketing discourse by engaging directly in discussions on ready-made topics.

In the professional practice dimension, interdiscursive performances are shown in the appropriation of professional practices from various fields through four kinds of interdiscursivity. In order to make marketing discourse less promotional in appearance, mixed and embedded interdiscursivity are often adopted in circumstances about festivals-related or heated issues to arouse readers' interest. Switched interdiscursivity, on the other hand, takes the form of switching between product descriptions and comments or purchase links in specific product promotions. As for the chained interdiscursivity, it resorts to reports, interviews, profiles, and other kinds of professional practices to present a more vivid product promotion. Moreover, in the era of new media, Uniqlo makes full use of multi-mode resources such as pictures, videos and hyperlinks to help diversify and novelize its marketing activities, and forms large-scale marketing by cooperating with other network platforms such as Weibo, wechat and official website to expand its influence.

In the professional culture dimension, given their focus not only on the product itself but also on the services and other added values behind it, commercial activities always keep company with other professional cultures. The openness of Weibo further facilitates the integration of different cultures, so as to form three kinds of cultural tendencies through the mixture of four kinds of interdiscursivity. With the increasing convenience of human interaction, especially with the technological support of the Internet, social issues, Chinese traditions, marketing practices and other events are more easily linked together, resulting in more complex and diverse microblog discourse. In this way, readers' interest in the remaining parts is greatly stimulated, and the ultimate aim of promoting products has been achieved.

Références bibliographiques

Alafnan, M. A. 2017. Critical perspective to genre analysis: Intertextuality and interdiscursivity in electronic mail communication. Advances in Journalism and Communication, (1): 23-49.

Antoinette, F. B. 2015. Investigating interdiscursivity in hospital strategic plans using Foucauldian discourse analysis. Hermes, 27(1): 35-47.

Bhatia, A. 2017. Interdiscursive performance in digital professions: The case of YouTube tutorials. Journal of Pragmatics, 124: 106-120.

Bhatia, V. K. 2004. Worlds of Written Discourse: A Genre-based View. London: Continuum.

Bhatia, V. K. 2006. Discursive practices in disciplinary and professional contexts. Linguistic and Human Sciences, (1): 5-28.

Bhatia, V. K. 2010. Interdiscursivity in professional communication. Discourse and Communication, (1): 32-50.

Bhatia, V. K. 2017. Critical Genre Analysis: Investigating Interdiscursive Performance in Professional Practice. London: Routledge.

Compagnone, A. 2015. The reconceptualization of academic discourse as a professional practice in the digital age: A critical genre analysis of TED Talks. Hermes, (54): 49-69.

Deng, L., Laghari, T. & Gao, X. 2021. A genre-based exploration of intertextuality and interdiscursivity in advertorial discourse. English for Specific Purposes, (6): 30-42.

Liu, C. & Liu, H. 2022. Critical genre analysis of the "Discussion" section of MA Theses by international students. Journal of University of Science and Technology Beijing, (2): 172-178.

Liu, W., Han, Z., Chen, H., & Ren, W. 2021. Discursive practice of Chinese criminal adjudication: A genre perspective. Círculo de Lingüística Aplicada a la Comunicación, 86: 69-80.

Lung, J. 2015. Interdiscursivity in public relations communication: Appropriation of genre and genre resources. Hermes, (1): 21-33.

Qian, Y. 2020. A critical genre analysis of MD & A discourse in corporate annual reports. Discourse & Communication, (4): 424-437.

Rajandran, K. 2018. Coercive, mimetic and normative: Interdiscursivity in Malaysian CSR reports. Discourse & Communication, (4): 424-444.

Ren, W., Bhatia, V. K., & Han, Z. 2020. Analyzing interdiscursivity in legal genres: The case of Chinese lawyers' written opinions. Pragmatics and Society, (4): 615-639.

Sernovitz, A. 2006. Word of Mouth Marketing: How Smart Companies Get People Talking. New York: Kaplan Publishing.

Wang, Q. 2020. The new-media communication of the medical scientific popularization discourse -- From CGA perspective. Journal of University of Science and Technology Beijing, (5): 17-24.

Wu, J. 2012. Review of the studies on interdiscursivity. Foreign Languages and Their Teaching, (2): 17-22.

Wu, J., Chen, C. & Zheng, R. 2018a. Interdiscursivity and critical genre analysis of company profiles: A case study of the Chinese mainland companies from Fortune Global 500. Linguistic Research, (2): 172-184.

Wu, J., Niu, Z. & Huang, Z. 2018b. A critical genre analysis of advertising discourse in WeChat official accounts: A case study of three foreign supermarkets. Shandong Foreign Language Teaching, (4): 30-37.

Yang, L. & Huang, Y. 2021. A critical genre analysis of product introduction discourse in cross-border E- Commerce -- A case study of 3C product introduction discourse. Journal of the Open University of Guangdong, (2): 64-70.

Apprenants sinophones du français et formes passives dans les écrits académiques

Wuran Lin [1], Marie-Paule Jacques [1]
[1] Laboratoire LIDILEM, Université Grenoble Alpes
wuran.lin@univ-grenoble-alpes.fr, marie-paule.jacques@univ-grenoble-alpes.fr

Introduction

Notre travail se focalise sur les emplois du passif des apprenants sinophones du FLE dans la rédaction de textes académiques en français, dans des disciplines de Lettres et Sciences Humaines. Nous nous interrogeons sur la façon dont les formes passives sont employées par ces apprenants. Pour cerner ces emplois, nous les comparons avec ceux des étudiants natifs et avec ceux de chercheurs francophones en textes scientifiques de disciplines des mêmes champs. Si des différences pour les emplois du passif sont observées, nous cherchons à en identifier les raisons.

Notre communication explicitera notre analyse fondée sur différents corpus que nous décrivons plus loin. Elle livre aussi les résultats d'une première étude.

De nombreuses études sur les apprenants sinophones ont abordé cette question pour l'anglais langue seconde (Chen, 2002 ; Liu & Feng, 2010 ; Wu & Que, 2013 ; Lin & al., 2017) mais peu pour le français langue étrangère. C'est pour cette raison que nous voulons documenter ces emplois. La question est toutefois mentionnée dans des travaux s'intéressant aux erreurs des apprenants sinophones du français (Zhang, 2016, dans son test de français[79]; Li, 2021, pour les textes argumentatifs), ainsi que dans d'autres travaux qui adoptent une perspective lexicale (Yan, 2017, pour les textes académiques). Notre intérêt est suscité par les difficultés particulières pour le passif des apprenants sinophones, montrées par ces travaux. Pour faire une analyse la plus complète possible des diverses formes passives écrites, nous avons choisi un contexte textuel dans lequel il existe par nature un grand nombre d'occurrences variées à étudier.

Notre choix s'est donc porté sur les textes scientifiques où les formes passives sont plus abondantes que dans d'autres types de textes, par exemple les textes journalistiques et littéraires (Fifielska, 2015) ou les discours oraux (Hamma et al., 2017). En outre, les apprenants sinophones doivent maitriser les diverses formes des écrits académiques qui se rapprochent des textes scientifiques.

[79] Le petit test de français de Zhang (2016) comprend deux parties. La première partie est la transformation des phrases actives en phrases passives. La deuxième partie est la traduction en français des phrases chinoises de sens passif.

Corpus et méthodologie

On ne peut dégager les spécificités des emplois du passif par des apprenants sinophones seulement d'une analyse directe des textes. C'est pourquoi, pour les mettre en évidence, nous procèderons à diverses comparaisons qui sont explicitées maintenant. Nos analyses se situent dans le champ des sciences humaines et de la littérature. Nous contrastons des écrits académiques en français d'étudiants sinophones avec des écrits de même genre d'étudiants natifs. De même, nous comparons des textes scientifiques de chercheurs natifs francophones et de chercheurs sinophones. Cette dernière comparaison fournira la référence des emplois pour les « experts » à partir de laquelle nous pourrons évaluer les emplois des étudiants.

En effet, pour proposer une description linguistique des formes passives employées dans l'écrit scientifique en français, qui serve de référence pour l'analyse des emplois des étudiants, nous avons besoin de textes rédigés par des experts natifs. Le corpus Scientext[80] propose un tel corpus, nous y avons sélectionné 22 articles scientifiques dans plusieurs disciplines des Sciences Humaines.

De la même manière, pour cerner les emplois du passif dans les textes scientifiques en mandarin et pour mettre en évidence d'éventuelles interférences entre la langue maternelle et la langue seconde, un corpus d'articles rédigés par des chercheurs sinophones est constitué de 80 articles scientifiques, concernant plusieurs disciplines de Sciences Humaines, issus du Lancaster Corpus of Mandarin Chinese version 2 (LCMCv2)[81]. Le corpus de chercheurs francophones et le corpus de chercheurs sinophones ont le même type et genre de textes, et sont de même rédigés par des natifs, et dans des disciplines similaires.

Pour pouvoir mieux cerner ce qui tient pour les étudiants aux difficultés linguistiques et distinguer ce qui relève de l'apprentissage d'un genre textuel nouveau, nous avons besoin d'écrits académiques produits par des étudiants natifs. Le corpus Littéracie Avancée[82] comporte 75 mémoires de master d'étudiants francophones de diverses disciplines et universités (Jacques et Rinck, 2017). Ces écrits sont, entre autres, destinés à permettre aux étudiants de s'acculturer à l'écriture scientifique. Dans l'étude préliminaire que nous présentons ici, nous avons exploité 15 mémoires pour comparaison avec ceux des étudiants sinophones. Ils nous permettent de mettre en lumière les difficultés que même les étudiants natifs affrontent.

Nous avons mené une première étude de l'écrit des apprenants sinophones sur un corpus composé de 15 mémoires issus du corpus de Rui Yan (2017). Ils sont écrits par des étudiants chinois en master de spécialité de français à l'Université des Études Internationales de Xi'an. Ces apprenants sinophones apprennent le français depuis six ans en moyenne et atteignent un niveau B2 ou C1 du CECRL (Cadre européen commun de référence pour les langues). Ces étudiants sont alors supposés maitriser les formes passives en français. Nous évoquons plus loin quelques résultats obtenus par la comparaison des formes passives de ce corpus avec celles des trois autres corpus.

---

[80] Le corpus Scientext est établi par Tutin et Grossmann (2014), consultable via https://corpora.aiakide.net/scientext20/?do=SQ.setView&view=corpora

[81] Le corpus LCMCv2 est élaboré par Richard Xiao, disponible via http://114.251.154.212/cqp/lcmc2/index.php?thisQ=search&uT=y

[82] Le corpus Littéracie Avancée est élaboré par F. Rinck, F. Boch, et M.-P. Jacques, , disponible via https://www.ortolang.fr/market/corpora/litteracieavancee

L'exploitation de ces quatre corpus s'appuie sur trois outils d'exploration textuelle : ScienQuest et Nooj pour le français, CQPweb pour le chinois (corpus de chercheurs). Ces outils nous ont permis de relever toutes les formes passives grâce aux requêtes. Les deux corpus d'étudiants ont été annotés au plan textuel en format XML. Nous avons eu recours à Nooj[83] puisqu'il permet de traiter automatiquement les corpus aux niveaux morphologique et syntaxique et d'extraire les constructions passives à l'aide de grammaires.

À partir de ces différentes extractions, nous avons procédé à diverses comparaisons dont nous indiquons maintenant quelques résultats.

Résultats

Les emplois du passif des apprenants sinophones ne sont pas identiques à ceux des étudiants natifs pour trois aspects analysés : la fréquence, la présence d'un complément d'agent et le verbe utilisé. Par rapport aux étudiants natifs, les formes passives sont sous-employées par les apprenants sinophones. Il s'agit d'un phénomène général qui est observé pour toutes les formes du passif (aussi bien V + PP, que passif réflexif ou autre). On note par ailleurs que dans les corpus d'écrits scientifiques, les formes passives sont moins fréquentes dans le corpus de chinois que dans le corpus de français. En effet, l'analyse des corpus de chercheurs montre que l'écrit scientifique en chinois utilise moins le passif que l'écrit scientifique en français.

Autre différence, la fréquence de l'expression du complément d'agent est largement plus élevée chez les apprenants sinophones que chez les étudiants natifs. Cette préférence de l'accompagnement du complément d'agent concerne elle aussi toutes les formes passives. Dans les corpus de chercheurs, les sinophones expriment plus fréquemment le complément d'agent que les francophones. On note aussi que les apprenants utilisent des pronoms personnels toniques en tant que complément d'agent. Ce phénomène est fréquent en chinois mais est présenté comme n'étant pas à préférer en français (Delatour et al., 2004).

De plus, la plupart des verbes les plus utilisés au passif par les apprenants sinophones diffèrent de ceux des étudiants et des chercheurs francophones. Les verbes les plus utilisés au passif par les apprenants sinophones sont analogues à ceux qu'on trouve dans les productions des chercheurs sinophones natifs. Cette comparaison montre que le choix des verbes au passif en français est influencé pour les apprenants sinophones par les valeurs pragmatiques du passif de leur langue maternelle.

En ce qui concerne les emplois du passif chez les francophones natifs, les étudiants utilisent moins fréquemment la forme passive que les chercheurs. Cela peut être expliqué par le fait que les étudiants natifs ne rencontrent pas de difficultés au niveau de la structure formelle du passif, mais au niveau des contextes d'emploi : ils ne sont pas conscients du fait que la construction passive soit une construction spécifique à l'écrit scientifique et académique.

Ces différentes comparaisons tendent à montrer une influence de la langue maternelle des étudiants sinophones sur leurs emplois du passif en français. Mais les emplois du passif des apprenants sinophones peuvent être influencés par d'autres facteurs, tels que le style personnel d'écriture, la maîtrise insuffisante des constructions syntaxiques, l'influence de l'anglais, etc. Notre travail se poursuivra avec des corpus plus volumineux qui mobiliseront des analyses quantitatives et qualitatives pour identifier de façon plus fine les différents facteurs en jeu pour

[83] Nooj est téléchargeable gratuitement sur le site : http://www.nooj4nlp.net.

l'emploi des passifs. Notre perspective à long terme est de fonder sur nos analyses linguistiques des pistes pédagogiques pour un enseignement du FLE qui tienne compte des spécificités des apprenants.

Références bibliographiques

Chen, W.X. 陈万霞 (2002). 从中国学习者英语语料库看英语被动语态习得 (L'acquisition de la voix passive en anglais à travers du corpus anglais des apprenants chinois). 外语教学与研究 (*Enseignement et recherche en langues étrangères*), 5, 198-202.

Delatour, Y., Jennepin, D., Léon-dufour, M., & Teyssier, B. (2004). *Nouvelle grammaire du français. Cours de civilisation française de la Sorbonne.* Paris : Hachette FLE.

Fifielska, E. (2015). *Les constructions syntaxiques de l'écrit scientifique : Exploration et analyses de corpus* [Mémoire de master]. Université Grenoble Alpes.

Hamma, B., Tardif, A. et Badin, F. (2017). Le passif à l'oral. In P. Larrivée et F. Lefeuvre (éd.), *Français contemporain vernaculaire (FRACOV)*. 1-15. En ligne à l'adresse suivante: http://www.univ-paris3.fr/medias/fichier/passif-a-l-oral_1486478858794.pdf.

Jacques, M.-P., & Rinck, F. (2017). Un corpus de "littéracie avancée" : Résultat et point de départ. *Corpus*, 16, 217-237.

Li, Q.Y. (2019). *Étude de la linéarité et des enchaînements dans les productions écrites d'apprenants sinophones du niveau avancé en français langue étrangère.* [Mémoire de master]. Université Grenoble Alpes.

Lin, H., Wu, M.F., & Feng, Y. (2017). 基于语料库的国内英语学习者被动语态使用的对比分析 (L'analyse comparative de l'utilisation de la voix passive basée sur un corpus par des apprenants de l'anglais). 沈阳建筑大学学报 (*Journal of Shenyang Jianzhu University*), 19 (5), 524-529.

Liu, J.W & Feng, Z.X. 刘敬伟&冯宗祥(2010). 我国英语专业研究生学位论文被动语态的使用分析 (L'analyse de l'utilisation de la voix passive dans les mémoires des apprenants chinois en master de spécialité de l'anglais). 沈阳农业大学学报 (J*ournal of Shenyang Agricultural University*), 12, 327-329.

Wu, Y. & Que, Z.J. 吴颖 & 阙紫江 (2013). 基于语料库的中国学生议论文写作中的被动语态习得研究 (Étude du corpus sur l'acquisition de la voix passive dans les textes argumentatifs des apprenants chinois), 外语教育 (*Enseignement des langues étrangères*), 107-114.

Yan, R. (2017). *Étude des constructions verbales scientifiques dans une perspective didactique: utilisation des corpus dans le diagnostic des besoins langagiers en FLE à l'aide des techniques de TAL.* Thèse: Linguistique. Grenoble: Université Grenoble Alpes.

Zhang, L. (2016). *Analyse des difficultés rencontrées par les étudiants chinois au cours de leur apprentissage du français et réflexions didactique.* Thèse: Linguistique. Wuhan : Université de Wuhan.

Mesurer l'accord inter-juge avec l'Alpha de Krippendorff : une étude des fonctions de différence

Jonas Noblet [1]
[1] Laboratoire LIDILEM, Université Grenoble Alpes
jonas.noblet@univ-grenoble-alpes.fr

Résumé

De nombreux événements (prise de parole, mouvement social, réaction chimique...) disparaissent tout juste après avoir été observés. Il est donc nécessaire de consigner les caractéristiques de ces événements pour témoigner qu'ils ont existé, permettre leur analyse et pour conserver et transmettre ces informations. Pour qu'un témoignage — les données recueillies — soit interprétable, il est nécessaire de pouvoir faire confiance à celui qui l'a consigné. En pratique dans la recherche, la confiance se traduit généralement sous la forme de *reproductibilité* : la probabilité que pour deux objets ou événements identiques la prise de données soit la même, indépendamment des circonstances propres à la collecte (Krippendorff, 2018).

L'une des manières de recueillir des données est d'effectuer une mesure : on emploie le terme mesure pour désigner « l'assignation d'un numéro (qu'on appelle étiquette) à un objet ou un événement selon un système de règles » (Stevens, 1946). Une mesure peut aussi bien être l'enregistrement de la température de l'air (mesure à échelle d'intervalle), que le numéro donné à un joueur de hockey (mesure à échelle nominale). La classification de données linguistiques, en considérant que chaque classe peut être représentée sous forme numérique, est ainsi un cas spécifique de mesure.

L'aspect reproductible (ou la *fiabilité*, en anglais *reliability*) d'une mesure est l'un des deux critères principaux, avec la validité, qui permet de lui donner du sens (Krippendorff, 2018). L'une des manières de tester la fiabilité est d'évaluer, pour une même mesure, la capacité d'accord de plusieurs annotateurs — les individus ou outils qui effectuent la mesure. En d'autres termes, est testée la capacité qu'ont les annotateurs à assigner des étiquettes similaires à des objets similaires.

Plusieurs méthodes de calcul ont été proposées pour quantifier le degré d'accord entre annotateurs. L'une des plus employées, malgré certains défauts (Feinstein et Cicchetti, 1990) et des restrictions sur son utilisation, est le Kappa (κ) de Cohen (Cohen, 1960). Des variations ont été proposées pour répondre à ses défauts ou limitations, comme le Kappa de Fleiss qui permet de prendre en compte trois ou plus annotateurs, ou le Kappa pondéré (Fleiss et al., 1969). D'autres méthodes ont également été développées comme le coefficient de Brennan-Prediger (Brennan et Prediger, 1981), le coefficient AC1/AC2 de Gwet (Gwet, 2008) ou plus récemment la méthode Gamma (Mathet et al., 2015).

Pour cette communication, nous proposons d'examiner une mesure d'accord spécifique, l'Alpha de Krippendorff (Krippendorff, 2018). L'Alpha de Krippendorff, contrairement à

d'autres mesures de l'accord inter-annotateurs, a été pensé pour répondre à des situations diverses. L'Alpha est en effet applicable quel que soit le nombre d'annotateurs et permet de prendre en compte plusieurs contraintes sur les données :

- données évaluées sur différents types d'échelle (échelle nominale, ordinale, d'intervalle...) ;
- échantillons de données de taille variable ;
- données avec des annotations manquantes.

Nous présenterons à cet égard les caractéristiques de l'Alpha en nous basant sur la méthode de calcul décrite par K. Krippendorff (Krippendorff, 2011). L'objectif de cette étude est la mise en valeur de certaines propriétés fondamentales et des possibles biais qui peuvent émerger. Plusieurs travaux se sont penchés sur la problématique de la modélisation de l'accord fortuit, par chance, dans les mesures d'accord (Zhao et al., 2013, Mathet et Widlöcher, 2016). Notre communication porte quant à elle sur la prise en compte des différentes échelles de mesures (échelles nominales, échelles d'intervalles). En particulier, nous nous intéresserons aux fonctions de différence qui jouent un rôle important dans le calcul de l'alpha.

Notre analyse se basera avant tout sur des considérations mathématiques. Nous examinerons les expressions des fonctions de différences pour expliquer le comportement de l'Alpha dans différents contextes.

Pour illustrer le propos, nous prendrons comme premier exemple une mesure ternaire (trois étiquettes) inspirée d'un projet de recherche en cours. Bien que formellement simple, cette mesure nous permettra de mettre en lumière certaines problématiques liées à l'estimation de l'accord inter- annotateur. Notamment, on relèvera des spécificités propres au codage qualitatif de données linguistiques.

Dans le cadre de la mesure ternaire, nous proposerons plusieurs scénarios d'annotations, qui permettront de corroborer la réflexion mathématique. Nous étendrons ensuite l'analyse à un ensemble plus large de mesures.

La finalité de l'étude sera d'avancer un ensemble de recommandations pour guider l'interprétation de l'Alpha.

Références bibliographiques

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa : Some Uses, Misuses, and Alternatives [Publisher : SAGE Publications Inc]. *Educational and Psychological Measurement*, *41* (3), 687- 699. https://doi.org/10.1177/001316448104100307

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales [Publisher : SAGE Publications Inc]. *Educational and Psychological Measurement*, *20* (1), 37-46. https :// doi .org/10.1177/001316446002000104

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa : I. the problems of two paradoxes [Publisher : Elsevier]. *Journal of Clinical Epidemiology*, *43* (6), 543-549. https://doi.org/10.1016/0895-4356(90)90158-L

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa [Place : US Publisher : American Psychological Association]. *Psychological Bulletin*, *72*, 323-327. https://doi.org/10.1037/h0028106

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agree- ment [_eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1348/000711006X126600]. *British Journal of Mathematical and Statistical Psychology*, *61* (1), 29-48. https://doi.org/10.1348/ 000711006X126600

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. *Departmental Papers (ASC)*. https://repository.upenn.edu/asc_papers/43

Krippendorff, K. (2018). *Content Analysis : An Introduction to Its Methodology*. SAGE Publica- tions.

Mathet, Y., & Widlöcher, A. (2016). Évaluation des annotations : Ses principes et ses pièges. *TAL Traitement Automatique des Langues*, *57*, 73-98.

Mathet, Y., Widlöcher, A., & Metivier, J.-P. (2015). The Unified and Holistic Method Gamma for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, *41*. https://doi.org/10.1162/COLI_a_00227

Stevens, S. S. (1946). On the Theory of Scales of Measurement [Publisher : American Association for the Advancement of Science]. *Science*, *103* (2684), 677-680. Récupérée 6 janvier 2023, à partir de https://www.jstor.org/stable/1671815

Zhao, X., Liu, J., & Deng, K. (2013). Assumptions behind Intercoder Reliability Indices. *Annals of the International Communication Association*, *36*. https://doi.org/10.1080/23808985.2013. 11679142

Les rédactions des étudiants : constitution d'un corpus d'erreurs syntaxiques

Laura Noreskal[1], Iris Eshkol-Taravella[1] et Marianne Desmets[2]
[1] MoDyCo UMR-7114, Université Paris Nanterre
[2] LLF UMR-7110, Université Paris Nanterre
laura.noreskal@parisnanterre.fr, ieshkolt@parisnanterre.fr, marianne.desmets@parisnanterre.fr

Introduction

De nos jours, de nombreux étudiants rencontrent des difficultés rédactionnelles lors de leur entrée dans l'enseignement supérieur. Cependant, les universités n'ont pas toujours les outils nécessaires pour aider les étudiants et se retrouvent souvent dans l'obligation de développer localement des solutions pédagogiques. C'est face à ce constat qu'une quarantaine d'universités et d'institutions ont décidé de se réunir pour proposer des solutions de remédiation dans le cadre du projet national écri+[84]. Le projet écri+ a pour objectif de proposer des outils d'évaluation, de formation et de certification pour l'amélioration des compétences langagières des étudiants francophones. De cette collaboration est né un référentiel de compétences rédactionnelles à maîtriser. Parmi les difficultés observées, les constructions syntaxiques complexes et les séquences phrastiques longues, avec des coordinations ou des juxtapositions sont apparues comme posant particulièrement problème aux étudiants. Ainsi, dans un objectif d'auto-formation des étudiants sur la reconnaissance de ce type d'erreurs dans leurs textes, le projet propose de mettre à leur disposition un outil de détection automatique d'erreurs dans les phrases coordonnées. Dans le cadre du projet, notre objectif est donc d'étudier les erreurs syntaxiques présentes dans les structures coordonnées ainsi que les différents critères qui les font apparaître dans les rédactions des étudiants. Pour ce faire, nous avons émis des hypothèses que nous avons vérifiées par une approche fondée sur la linguistique de corpus.

Objet étudié

Dans le cadre de cette recherche, l'objet étudié est la structure coordonnée erronée. Cependant, afin d'étudier la structure erronée, il est nécessaire de caractériser la structure coordonnée correcte. Ainsi, en se basant sur les définitions de Martinet (1980), de Goosse et al. (2008), de Riegel et al. (2009) et d'Abeillé et Godard (2021), il a été possible de relever deux faits importants sur la coordination :

— la coordination explicite requiert un coordonnant alors que la coordination implicite n'en attend pas, on parle alors de juxtaposition ;

— le rôle ou statut des éléments conjoints est important car il doit être respecté à chaque ajout.

Pour repérer les constructions coordonnées dans les rédactions des étudiants, nous avons donc décidé d'étudier celles qui contiennent de la coordination explicite, soit des conjonctions ou

[84] anr-17-NCUN-00015

des adverbes de liaison. Nous appelons donc structures coordonnées les constructions qui contiennent une conjonction de coordination (ou, et, or, ni, car, soit. . . soit) ou un adverbe de liaison (puis, ensuite, cependant, néanmoins. . .) qui permet de relier des mots, des syntagmes ou des phrases dans un énoncé.

Typologie des erreurs

À partir des phrases coordonnées erronées du corpus, il a été possible de distinguer plusieurs types d'erreurs syntaxiques. De là, nous avons pu établir une typologie des erreurs comprenant trois grands types d'erreurs :

— les problèmes de prépositions : les prépositions sont souvent sujettes aux erreurs dans les productions étudiantes. Trois sous-types d'erreurs sont distingués : l'absence de préposition (1), l'ajout d'une préposition inattendue (2) et le remplacement d'une préposition par une autre (3).

- Absence d'une préposition (PREP ABS)

  (1) *Ils illustrent leur propos en appliquant cette analyse **aux** appendices et **les** formules illocutoires, **les** actes indirects et **les** questions biaisées.

- Ajout d'une préposition non-attendue (PREP ADD)

  (2) *Avant de visionner la comédie musicale, il faudra étudier avec les élèves la période révolutionnaire pour comprendre les raisons de la Révolution et **de** rendre cette activité ludique mais pédagogique.

- Remplacement d'une préposition par une autre (PRED REMP)

  (3) *Le fait de les aider à se construire et **à** les voir grandir, tout en leur apportant un savoir doit être gratifiant et réjouissant.

— les problèmes entre conjoints : les groupes coordonnés doivent respecter des contraintes de parallélisme afin d'assurer la compréhension du texte. Il arrive cependant que ces contraintes ne soient pas respectées et qu'elles aient un impact sur la compréhension et la lisibilité de la coordination. Dans les rédactions des étudiantes, nous avons été confrontées à deux problèmes liés aux groupes conjoints : la mauvaise cohérence des groupes syntaxiques et grande distance entre conjoints.

- Mauvaise cohérence des groupes syntaxiques (MCGS)

  (4) *Par ailleurs, il appartient à tous les personnels de transmettre aux élèves les valeurs **et doivent** avoir un devoir de neutralité.

- Grande distance entre conjoints (DIST CONJOINT)

  (5) *S'associer avec des associations telles que celles du Téléthon qui permet de sensibiliser les élèves, de transmettre des valeurs républicaines, **ainsi que Nettoyons la Nature**.

— les erreurs d'accords sujet-verbe : les problèmes d'accords sont courants dans les productions universitaires. Nous avons donc décidé de prendre en compte les accords entre le sujet et le verbe car ils impliquent des liens entre les groupes de mots.

- Mauvais accord sujet-verbe (MASV)

  (6) *Le personnage de droite est assis sur un tabouret et **as** une corpulence fine.

Hypothèses

Pour répondre aux enjeux de cette recherche, un certain nombre d'hypothèses ont été formulées. Nous avons notamment été amenées à nous questionner sur les critères qui pouvaient influencer l'apparition des erreurs. Ainsi, nous avons cherché à savoir si le type de rédactions pouvait influencer la présence d'erreurs. De même, en ce qui concerne les constructions coordonnées erronées, il fallait savoir si la longueur des phrases et le nombre de coordonnants dans une phrase pouvaient également être des éléments discriminants.

Le type de la rédaction

Afin de vérifier si le type de la rédaction pouvait influencer l'apparition d'erreurs, un corpus permettant d'étudier les erreurs dans les rédactions des étudiants a été constitué. Pour ce faire, nous avons collecté des productions dites « évaluatives », c'est-à-dire des productions réalisées dans le but d'être évaluées par un enseignant dans le cadre d'un enseignement supérieur. Quatre types de rédactions ont été retenues : les devoirs maison, les exercices faits en classe, les rapports de stage et les mémoires. Le corpus de rédactions est donc composé de 380 rédactions comprenant 139 devoirs maison, 167 exercices, 47 rapports de stage et 27 mémoires. Les rédactions collectées étaient sous forme dactylographiée et ont été anonymisées. Cependant, trois informations ont été gardées en tant que métadonnées : l'année d'étude (L1, L2, L3, M1 et M2), le domaine d'étude (principalement Sciences du Langage, Sciences de l'éducation, droit et histoire) et le type de la rédaction (DM, exercices, rapports de stage et mémoires) afin de pouvoir faire des analyses sociolinguistiques. Par la suite, des phrases coordonnées correctes et erronées ont été extraites à partir des écrits collectés, en utilisant des patrons morphosyntaxiques créés avec l'outil Unitex (Paumier, 2011) qui reconnaissent les conjonctions de coordination et les adverbes de liaison. Après observation du corpus de phrases coordonnées, nous avons réalisé que les erreurs étaient plus nombreuses dans les phrases extraites des exercices avec une fréquence relative de nombre d'erreurs par nombre de phrases de 58% soit environ une erreur toutes les deux phrases.

Types	Nb Phrases correctes	Nb Phrases erronée	Total phrases	Fréquence relative nb erreur/nb phrase	1 erreur toutes les x phrases environ
Devoirs maison	806	515	1321	39%	2
Exercices	204	283	487	58%	3
Mémoires	447	177	624	28%	4
Rapports	445	193	638	30%	4

table 6. : **Distribution des phrases en Erronée/Correcte par type de rédactions**

Afin d'avoir une analyse approfondie, nous avons décidé de faire le test $\chi 2$ d'indépendance. Ce test consiste à vérifier s'il existe une relation de dépendance entre deux variables pour une population donnée. Le test $\chi 2$ est très utilisé en linguistique de corpus car il permet de mettre en évidence certaines variations linguistiques (Leech et Fallon, 1992 ; Hou *et al.*, 2021). Nous avons donc voulu savoir si les variables types (DM, exercices, mémoires et rapports) et erreurs

sont dépendantes l'une de l'autre. Nous avons donc émis l'hypothèse que l'apparition d'erreurs est liée aux types de rédactions. Afin de vérifier cette hypothèse, nous avons constitué un tableau de contingence avec les valeurs obtenues et les valeurs théoriques. Les valeurs théoriques permettent de savoir comment seraient réparties les données, s'il n'y avait pas de dépendance entre les variables.

Types		Correctes	Erronées	Total
Devoirs maison	Obtenu	806	515	1321
	Attendu	818	503	
Exercices	Obtenu	204	283	487
	Attendu	302	185	
Mémoires	Obtenu	447	177	624
	Attendu	387	237	
Rapports	Obtenu	445	193	638
	Attendu	395	243	

table 7. : **Tableau de contingence valeurs théoriques par types de rédactions**

Pour vérifier l'hypothèse, nous avons calculé la valeur de $\chi 2$ pour nos variables ainsi que le degré de liberté. Nous avons obtenu un degré de liberté de 3 et une valeur de 124,93. Nous avons choisi d'avoir une marge d'erreur de 5%. En comparant la valeur obtenue avec celle de la table de la loi du $\chi 2$ (7,815), nous remarquons que 124,93 est supérieur à 7,815, ce qui signifie que notre valeur statistique est supérieure à la valeur de la table de la loi du $\chi 2$ malgré le faible pourcentage (5%) de chances d'être supérieur. Nous pouvons donc confirmer l'hypothèse car il existe bien une relation entre les deux variables.

La longueur des phrases et le nombre de coordonnants

Concernant les phrases coordonnées, nous voulions savoir si la longueur d'une phrase et son nombre de coordonnants pouvaient être reliés à la présence d'une erreur dans la phrase. Pour vérifier ces hypothèses, nous avons décidé d'observer les moyennes, les minimum, les maximum et les écart-types du nombre de mots et de coordonnants par phrase.

	Nombre de mots		Nombre de coordonnants	
	Correctes	Erronées	Correctes	Erronées
Moyenne	21,1	28,5	1.383	1.657
Minimum	4	6	1	1
Maximum	104	95	7	8
Ecart-type	9,6	13	0.679	0.948

table 8. : **Comparaison du nombre de mots et de coordonnants dans les phrases correctes et les phrases erronées**

Type d'erreur	Moyenne de mots/phrase	Moyenne de coordonnants/phrase
DIST CONJOINT	30,05	1,89
MCGS	25,34	1,68
PREP ABS	25,89	1,69
PREP ADD	33,04	2
PREP REMP	26,10	1,38
MASV	25,12	1,5

table 9. : **Moyenne du nombre de mots et du nombre de coordonnants par phrase pour chaque type d'erreur**

En premier lieu, nous pensions que le nombre de mots dans une phrase pouvait influencer la présence ou non d'erreurs. En observant les données, nous avons pu remarquer que les

phrases erronées étaient en moyenne plus longues que les phrases correctes. Afin de mieux comprendre ces chiffres, nous avons observé les phrases erronées les plus longues. Nous avons pu remarquer que les types d'erreurs les plus présents dans les phrases longues étaient les erreurs d'ajout de préposition et les erreurs de distance entre conjoints. Dans un second temps, nous avons voulu vérifier l'influence du nombre de coordonnants sur la présence d'erreurs. Le nombre de coordonnants était plus élevé dans les phrases erronées que dans les phrases correctes. Là encore, nous avons observé les phrases erronées comprenant le plus de coordonnants. Les erreurs d'ajout de préposition et de distance entre conjoints ont également été les types d'erreurs les plus présents dans les phrases longues. Ces observations nous ont alors permis d'observer que les deux types d'erreur ajout de préposition et distance entre conjoints sont liés aux nombres de mots et de coordonnants dans la phrase. Cependant, cela n'a pas suffi à valider nos hypothèses de départ. En effet, d'autres tests statistiques pourraient être réalisé pour mieux analyser les différences entre les phrases erronées et les phrases correctes.

Références bibliographiques

Abeillé, A., Godard, D. (2021). La grande grammaire du français. Éditions Actes Sud.

Goosse, A., Grevisse, M.(2008). Le bon usage. De Boeck Superieur.

Landis, J. R., Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics,* p. 363-374

Martinet, A. (1980). Éléments de linguistique générale. *Collection U Prisme*. Albin Michel.

Paumier, S. (2011). Unitex-manuel d'utilisation.

Riegel, M., Pellat, J., Rioul, R. (2009). Grammaire méthodique du français [1994]. Paris : PUF

Leech, G. Fallon, R. (1992). Computer corpora-what do they tell us about culture. *ICAME journal 16*.

Hou, J., Landragin, F. (2021). L'effet des facteurs de distance et de fréquence sur la saillance des entités référentielles. *Langages* 224.4, p. 109-128

Quelle sémantique pour les verbes modaux du français ? Étude des propriétés combinatoires de *pouvoir*, *devoir*, *falloir* et *vouloir*

Aylin Pamuksaç[1]
[1] Chaire de Linguistique française, Institut des Sciences du Langage, Université de Neuchâtel
aylin.pamuksac@unine.ch

Introduction

La modalité est un concept qui a été défini de nombreuses fois par des autrices[85] venant de disciplines diversifiées comme la philosophie, la logique et la linguistique (Larreya 2004 : 733). Dans ce dernier domaine, le plus souvent, les études concernent la sémantique des verbes modaux, et plus précisément les effets de sens et les valeurs modales[86] de *pouvoir* et *devoir*. Elles s'inscrivent ainsi dans les conceptions étroites de la modalité et sont « centrées sur les notions de nécessaire et de possible » (Gosselin 2010 : 6). Ces deux verbes sont souvent comparés (Barbet 2012 ; De Saussure 2014 ; Hütsch 2020 ; Leeman-Bouix 2002 ; Sueur 1977, 1979 ; Vetters 2004, 2012) mais aussi étudiés de manière individuelle (cf. pour *devoir* : Bres (2022) ; Dendale (1994, 2000) ; Desclés et Guentchéva (2001) ; Kronning (1990, 1996, 2001) ; Marque-Pucheu (2001) ; Rossari et al. (2018) ; Vetters et Barbet (2006) ; cf. pour *pouvoir* : Le Querler (2001) ; Tasmowski et Dendale (1994)).

Pourtant, ce ne sont pas les seuls verbes modaux qui permettent de dénoter les modalités déontiques (du nécessaire) ou épistémiques (du possible). En effet, *falloir* et *vouloir* peuvent aussi exprimer une obligation (1 et 2), ou, dans une moindre mesure, une probabilité (3 et 4) :

1. Il *faut* ranger ta chambre.
2. Je *veux* que tu ranges ta chambre.
3. Il *faut* être bête pour croire que la terre est plate.
4. Il *veut* pleuvoir ce soir[87].

Falloir et *vouloir* font aussi l'objet d'études comparatives avec *pouvoir* et *devoir* même si elles sont moins nombreuses que pour ces derniers (pour *vouloir* comparé à *pouvoir* et

[85] Nous faisons le choix, dans ce travail, d'utiliser le féminin générique pour les noms mais aussi les pronoms. Ainsi, lorsqu'un groupe est composé de minimum une personne genrée au féminin, nous emploierons le pronom *elle·s*. Toutefois, lorsque cela s'avère possible, nous utiliserons les noms épicènes comme pour *professeur·e* et utiliserons *il·s* lorsque nous faisons référence à des personnes auxquelles sont attribuées le genre masculin.

[86] Par *effet de sens* nous entendons, par exemple, la *possibilité* et la *capacité* pour *pouvoir*. Nous qualifions de *valeur modale* les différentes modalités que les verbes modaux peuvent exprimer, comme les modalités déontiques ou épistémiques.

[87] Gadet (2021) considère cet usage de *vouloir* comme exprimant un futur périphrastique. Cependant, dans les exemples qu'elle soulève, soit « *Elle ne pourra pas étendre aujourd'hui parce que je sens que ça* ***veut*** *pleuvoir* » ou encore « *Descends de là, tu* ***veux*** *tomber* », ni la pluie, ni la chute ne semblent être des évènements certains d'arriver mais seulement potentiels et peuvent ainsi évoquer la modalité épistémique.

devoir : Desclés (2003) ; Larreya (1997) ; pour *falloir* comparé à *devoir* : Hooke (1935) ; Reed (2019) ; Rivière (1984) ; Valiukienė (2016) ; pour *falloir* comparé à *pouvoir* : Bryant (1980) ; pour une comparaison entre *pouvoir*, *devoir* et *falloir* : Caron et Caron-Pargue (2003)).

Malgré l'émergence des nouveaux outils numériques en linguistique, les verbes modaux font rarement l'objet d'études statistiques[88]. Dans cette recherche, nous désirons mesurer la ressemblance ou la divergence entre ces verbes modaux, qui ne sont pas seulement polysémiques – chacun des verbes mentionnés peut exprimer plusieurs effets de sens et valeurs modales – mais aussi synonymes – ils peuvent être interchangeables dans certains contextes comme le démontrent les exemples 1 à 4. Malgré cela, chacun de ces verbes semble pourtant posséder un certain noyau sémantique propre, comme la possibilité pour *pouvoir* (De Saussure 2014), la nécessité pour *devoir* (Vetters & Barbet 2006), la contrainte pour *falloir* (Caron & Caron-Pargue 2003) ou encore l'intention pour *vouloir* (Condoravdi & Lauer 2016)[89]. Ces proximités, éloignements mais aussi singularités, sont mesurés grâce aux propriétés combinatoires de ces quatre verbes modaux avec a) leur complémentation verbale sous forme de verbe à l'infinitif et b) les noms en position de sujet de ces verbes modaux.

Corpus et méthodologie

Corpus

Pour cette étude, nous effectuons nos recherches dans trois corpus[90] à visée informative :

I. *Le Monde* 2010, presse nationale française, 17'895'009 de tokens.
II. *L'Est Républicain* 2010, presse régionale française, 18'669'845 de tokens.
III. *Wikipédia* 2019, encyclopédie collaborative, 18'715'455 tokens tirés de 10'570 articles pris au hasard en 2019.

Il est intéressant d'investiguer des corpus à visée informative pour l'usage des verbes modaux car « ces sous-genres peuvent relater aussi bien des faits objectifs (comme dans les comptes rendus et reportages) que du jugement subjectif (comme dans les éditoriaux et commentaires) » (Hütsch 2020 : 15) et présenter ainsi une certaine diversité de la langue. Nous considérons nos corpus comme « un terrain d'étude » (Hütsch 2020 : 15) et adoptons ainsi une perspective de linguistique de corpus, en analysant « un marqueur linguistique décrit en usage » (Mayaffre et al. 2019 : 102).

[88] Kronning (1996) quantifie les proportions d'effets de sens de *devoir* dans divers corpus, Patard (2015) dénombre les fréquences relatives de la construction *faut croire (que)* dans des corpus issus de différentes tranches diachroniques, Valiukienė (2022) utilise des corpus parallèles pour calculer les effets de sens les plus communs de *falloir* à l'aide de traductions en lituanien et Blumenthal (2012) quantifie les fréquences relatives de *vouloir* et ses quasi-synonymes à travers les siècles dans Frantext.

[89] Condoravdi et Lauer (2016) accordent ce noyau sémantique de l'*intention* pour l'équivalent anglais *want* mais il nous semble applicable également à *vouloir* au vu des différents effets de sens qu'il peut véhiculer comme le *désir* (porté nécessairement sur le futur selon Blumenthal (2012)) et le *futur* (Riegel et al. 2008).

[90] Tous ces corpus ont été élaborés par nous-même ou des membres de l'équipe de la Chaire de Linguistique française de l'Université de Neuchâtel dont nous faisons actuellement partie. Ils ne forment pas un seul et même corpus mais trois corpus différents, tous investigués pour cette étude.

Méthodologie

Ces corpus sont exploités sur la plateforme TXM développée à l'ENS de Lyon (Heiden et al. 2010). Afin de distinguer ou rassembler les divers effets de sens et noyaux sémantiques des verbes modaux, nous nous intéressons aux combinaisons de ces verbes avec a) leur complémentation verbale sous forme de verbe à l'infinitif et b) les noms en position de sujet. Ces deux éléments sont intéressants à étudier, car ils font partie de la valence des verbes modaux. En effet, par leur nature de verbe, ils possèdent, tous un sujet (Lazard 2009 : 152-153) et ils « […] ne peuvent à eux seuls décrire un type de situation et doivent se combiner avec une autre forme verbale. » (Abeillé et al. 2021 : 131).

Ces éléments font l'objet de classifications sémantiques attestées comme la classification des verbes de Dubois et Dubois-Charlier (1997) et celle de noms de Salvadori et al. (2021)[91]. Ces deux typologies sont utilisées dans notre travail pour ventiler nos résultats dans des classes plus grandes que les formes lexicales seules, ce qui nous permet d'avoir une vision moins « dispersée » des nombreuses cooccurrences significatives des verbes modaux analysés.

Nous mesurons ainsi l'attractivité entre les verbes modaux et ces deux éléments – leur complémentation verbale sous forme de verbe à l'infinitif et les noms en position de sujet – grâce à l'indice de spécificité (Lafon 1980) mais aussi les cooccurrences brutes. Ces données nous aident à déterminer des profils sémantiques de ces quatre verbes en dehors des catégories modales prédéfinies.

Résultats

Les premiers résultats de notre recherche permettent de voir des différences fondamentales entre les verbes modaux. Par exemple, en ce qui concerne les sujets, nous pouvons remarquer que la distribution des types de noms est très différente d'un verbe à l'autre[92] :

[91] Il est évident que les verbes et les noms sont polysémiques et qu'ils appartiennent ainsi à plus d'une catégorie sémantique. Nous avons choisi de les insérer seulement dans une seule catégorie à chaque fois. Pour certains verbes et noms, la catégorie dominante était assez évidente par les diverses descriptions des catégories. Pour les cas plus discutables, nous avons sélectionné des échantillons de KWIC dans nos corpus et analysé la proportionnalité de représentation des diverses catégories. Nous avons ainsi inséré les verbes et les noms dans les catégories les plus fréquentes de nos échantillons.

[92] Dans cette figure, *falloir* est représenté à partir des noms qui figurent dans la complétive dans la construction *il FALLOIR que* ***NOM*** *VERBE*.

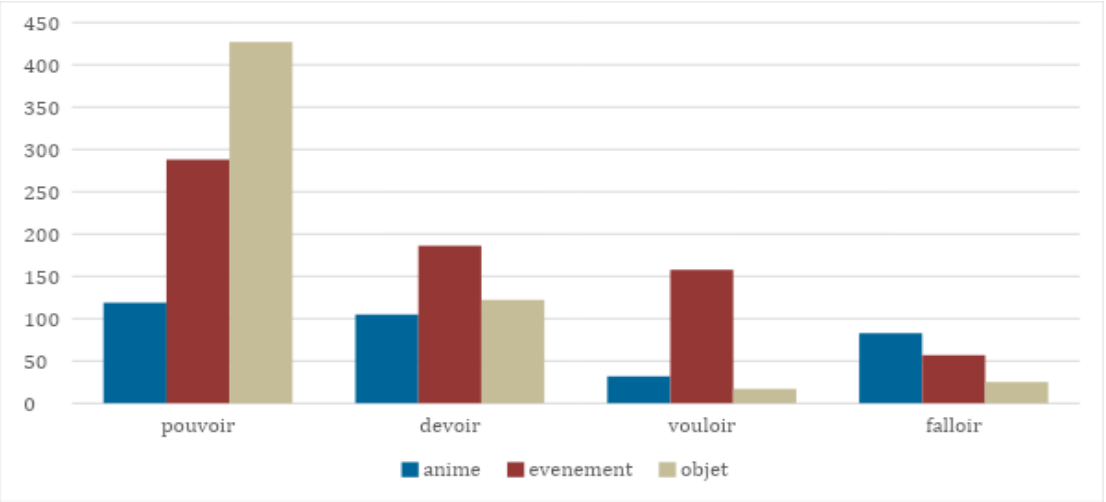


figure . 1 Classes des noms figurant en position sujet de *pouvoir*, *devoir*, *falloir* et *vouloir* en tant que lemme dans Wikipédia selon la classification de Salvadori et al. (2021) réduite à trois catégories dans Varvara et al. (2022). Les chiffres correspondent à des co-fréquences absolues.

Nous pouvons ainsi noter que *devoir* ne semble pas préférer un type de sujet plutôt qu'un autre même si la catégorie *évènement* est légèrement plus représentée. *Pouvoir*, quant à lui, préfère nettement les objets. Pour *falloir*, lorsque le sujet est exprimé dans la complétive, celui-ci prend, la plupart du temps, une forme 'animé'. Pour finir, *vouloir*, attire davantage les noms de type 'évènement'. Dans cette catégorie figure des noms comme *hypothèse* :

6. Une autre **hypothèse veut** que le hasard soit à l'origine de l'attribution des royaumes. (Wikipédia 2019)

Le noyau sémantique de l'*intention*, généralement attribué à un être animé[93], peut donc être rattachée à un nom plus abstrait. Ce résultat peut également faire écho à la grammaticalisation des formes *vouloir dire* dans lesquelles le verbe se désémantise :

7. En breton 'ro' signifie en effet 'donne', et 'sko' **veut dire** littéralement 'frappe', selon le contexte au sens physique de 'joue des poings'. (Wikipédia 2019)

Nos résultats révèlent aussi que l'association entre les verbes *dicendi* en position de complément n'est pas une particularité du verbe *vouloir*. *Pouvoir* et *falloir* sont également statistiquement associés à cette catégorie de verbes :

8. D'une manière générale, **on peut dire** que l'ordre des syntagmes est libre mais que la disposition des morphèmes à l'intérieur d'un syntagme est fixée par l'usage. (Wikipédia 2019)

9. **Il faut** également **noter** que qu'avec l'accroissement du nombre de locuteurs, l'espéranto est devenu la langue maternelle d'enfants issus de couple espérantophones. (Wikipédia 2019)

Pour *falloir*, Valiukienė (2022) avait déjà relevé que ce « verbe […] perd son poids sémantique de base dans des constructions qui sont plus ou moins figées aux niveaux lexico-sémantique et pragmatique. Sa connectivité avec des *verba cogitandi* et *dicendi*

[93] Car l'intention pousse à l'action (Deboës 2002 : 315), et ce sont généralement des être animés qui agissent.

entraîne sa dégrammaticalisation. » (Valiukienė 2022 : 11-12). Cependant, pour *pouvoir*, à notre connaissance, aucune étude ne relève cet emploi désémantisé du modal qui est pourtant très fréquent dans nos corpus.

Afin d'avoir une vue d'ensemble sur les similarités et disparités entre ces verbes, nous pouvons également associer les données de nos deux composantes et les mettre en lien à l'aide de l'analyse factorielle des correspondances (désormais AFC). A l'aide du logiciel Hyperbase (Brunet 2011) nous avons ainsi pu représenter les ressemblances ou divergences entre les quatre verbes modaux que nous étudions ici :

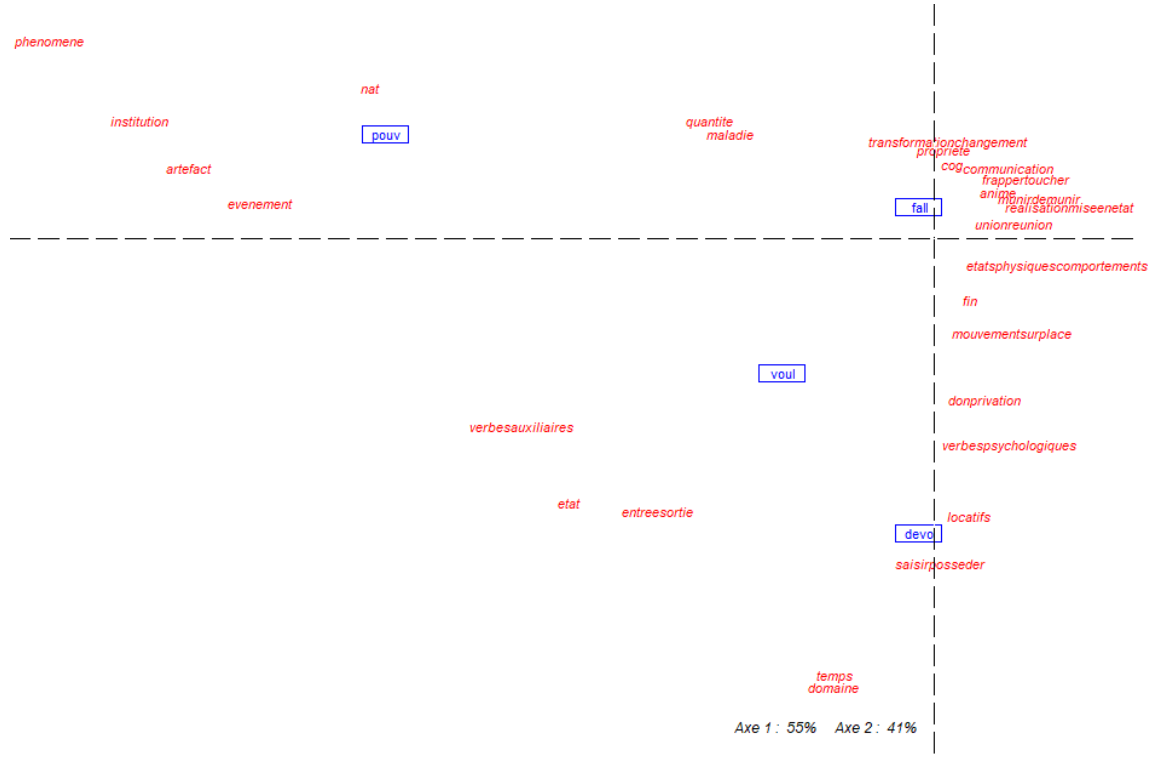


figure . 2 AFC des verbes modaux par lemme dans le corpus Wikipédia sur la base de leur co-fréquence avec les verbes en position de complémentation verbale (classés selon Dubois et Dubois-Charlier (1997)) et les noms en position sujet (classés selon Salvadori et al. (2021))

Dans la figure 2, nous pouvons ainsi observer que *pouvoir* semble se distinguer nettement de tous les autres verbes. Ceci peut être dû à son attraction à certains types de sujets qui semblent lui être propres comme les *artefacts* (objets créés par la Femme), les objets *naturels* (*nat* dans le graphique), ainsi que les noms de type *évènement* (qui correspondent généralement à des noms d'action ou de faits réalisées comme *annotation* ou encore *réaction*). C'est sur ce dernier élément que se joue l'attraction de *vouloir* sur la partie gauche du graphique et qui le rapproche ainsi de *pouvoir*. *Devoir* et *falloir* quant à eux sont très proches sur l'axe 1 (horizontal) mais se distingue sur l'axe 2 (vertical). Leurs différences semblent se jouer sur les *verbes psychologiques* (les verbes de pensée et les verbes d'affects) et les verbes locatifs (des actions sans déplacement comme *rester* mais aussi *placer*), ainsi que sur les sujets types 'domaine' (comme la *science* ou encore la *technique*), catégories majoritairement attirées par *devoir*[94]. Notons également que *falloir* est le verbe le plus proche du centre des axes ce qui

[94] Dans ce passage, nous nous sommes focalisées sur les catégories qui contribuent le plus aux axes dans notre AFC. Ces informations nous sont fournies par Hyperbase en pour mille.

démontre une certaine propriété « standard » de ce verbe qui ne se combine ainsi pas avec une ou plusieurs de ces catégories en particulier.

En mettant en parallèle ces résultats et ces diverses représentations des verbes modaux *pouvoir*, *devoir*, *falloir* et *vouloir* avec leur complémentation verbale sous forme d'un verbe à l'infinitif et les noms en position sujet, nous pouvons relever que

a) *vouloir* s'éloigne de son noyau sémantique en attribuant l'intention à des objets abstraits et en se grammaticalisant dans des locutions comme *vouloir dire*.
b) *pouvoir* et *falloir* partagent des propriétés en se désémantisant lorsqu'ils sont associés à des verbes *dicendi*.
c) *devoir*, quant à lui, ne semble pas faire l'objet d'un processus de grammaticalisation car ses cooccurrents spécifiques sont davantage diversifiés que pour les autres verbes modaux.
d) *pouvoir* et *vouloir* semblent aussi partager des sujets de type 'objets' (naturels ou non) ainsi que les noms 'évènement' et c'est ce qui les distinguent des deux autres verbes.
e) *falloir* et *devoir* partagent aussi beaucoup de propriétés communes mais se distinguent par rapport à des sujets et verbes spécifiquement associés à *devoir*.
f) *falloir* est le verbe modal le plus standard et peut ainsi se avec beaucoup de catégories sémantiques sans que l'une ou plusieurs d'entre elles ne se démarquent des autres.

Ces propriétés seront approfondies à l'aide des résultats d'autres corpus mais aussi en mettant en lumière d'autres cooccurrences significatives lors de la présentation.

Références bibliographiques

Abeillé, A., Koenig, J.-P., Godard, D., & Bonami, O. (2021). Qu'est-ce qu'un verbe ? In A. Abeillé, D. Godard, A. Delaveau, & A. Gautier (Eds.), *La Grande Grammaire du Français* (Vol. 1, pp. 127-147). Arles : Actes Sud.

Barbet, C. (2012). Devoir et pouvoir, des marqueurs modaux ou évidentiels? *Langue française*, 1, 49-63.

Blumenthal, P. (2012). Les implications de la volition: Types de procès, centrage du verbe et polyphonie. In M. Birkelund, G. Boysen, & P. S. Kierkegaard (Eds.), *Aspects de la Modalité* (Vol. 469, pp. 31-44). Tübingen : Max Niemeyer Verlag.

Bres, J. (2022). Devoir en emploi évidentiel reportif. *Langue française*, 215 (3), 43-60. https://doi.org/10.3917/lf.215.0043

Brunet, E. (2011). *Hyperbase, Logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels, Manuel de Référence*. Nice : Université de Nice.

Bryant, W. H. (1980). Unequivocal passé composé/imparfait Contexts for falloir and pouvoir. *The French Review*, 53 (4), 514-524. http://www.jstor.org/stable/391630

Caron, J., & Caron-Pargue, J. (2003). A multidimensional analysis of French modal verbs pouvoir, devoir and falloir. In F. H. van Eemeren, C. A. Willard, & F. Snoeck Henkemans

(Ed.) *Proceedings of the Fifth Conference of the International Society for the Study of Argumentation*, Amsterdam.

Condoravdi, C., & Lauer, S. (2016). Anankastic conditionals are just conditionals. *Semantics and Pragmatics*, 9, 1-69, Article 8. https://doi.org/10.3765/sp.9.8

De Saussure, L. (2014). Verbes modaux et enrichissement pragmatique. *Langages*, 193 (1), 113-126. https://doi.org/10.3917/lang.193.0113

Deboës, E. (2002). *Analyse sémantique des verbes français: aperçu d'une représentation du sens par des structures logico-sémantiques.* Thèse de doctorat. Lyon 2 : Lyon.

Dendale, P. (1994). 'DEVOIR' ÉPISTÉMIQUE, MARQUEUR MODAL OU ÉVIDENTIEL ? *Langue française*, 102, 24-40. http://www.jstor.org/stable/41559281

Dendale, P. (2000). Devoir épistémique à l'indicatif et au conditionnel: inférence ou prédiction? In A. Englebert (Ed.) *Actes du 22e Congrès international de Linguistique et de Philologie Romanes*, Bruxelles.

Desclés, J.-P. (2003). Interactions entre les valeurs de pouvoir, vouloir, devoir. *Aspects de la Modalité*, 469, 49-66.

Desclés, J.-P., & Guentchéva, Z. (2001). La notion d'abduction et le verbe devoir 'épistémique'. In *Les verbes modaux* (pp. 103-122). Leyde : Brill.

Dubois, J., & Dubois-Charlier, F. (1997). *Les Verbes français*. Paris : Larousse.

Gadet, F. (2021). La variation régionale des périphrases verbales. In A. Abeillé, D. Godard, A. Delaveau, & A. Gautier (Eds.), *La Grande Grammaire du Français* (Vol. 1, pp. 1273-1275). Arles : Actes Sud.

Gosselin, L. (2010). *La validation des Représentations : Les modalités en Français*. Amsterdam : Rodopi.

Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, Rome, Italy. https://halshs.archives-ouvertes.fr/halshs-00549779

Hooke, M. K. (1935). Some Neglected Uses of "Falloir" and "Devoir". *The Modern Language Journal*, 19 (4), 278-284. https://doi.org/10.2307/314714

Hütsch, A. (2020). *L'usage des verbes modaux en français et en allemand: étude contrastive de la combinatoire adverbiale sous l'éclairage quantitatif.* Thèse de doctorat. Université de Neuchâtel : Neuchâtel.

Kronning, H. (1990). Modalité et diachronie: du déontique à l'épistémique. L'évolution sémantique de debere/devoir. *Onzième Congrès des Romanistes Scandinaves, Trondheim 13-17 août 1990*, Trondheim.

Kronning, H. (1996). *Modalité, cognition et polysémie : sémantique du verbe modal devoir.* Thèse de doctorat. Acta Universitatis Upsaliensis : Uppsala.

Kronning, H. (2001). Pour une tripartition des emplois du modal devoir. In P. Dendale & J. Van der Auwera (Eds.), *Les verbes modaux* (pp. 67-84). Leyde : Brill.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1 (1), 127-165. https://www.persee.fr/doc/mots_0243-6450_1980_num_1_1_1008

Larreya, P. (1997). Notions et opérations modales: pouvoir, devoir, vouloir. In C. Rivière & M.-L. Groussier (Eds.), *La notion* (pp. 156-165). Paris : Ophrys.

Larreya, P. (2004). L'expression de la modalité en français et en anglais (domaine verbal). *Revue belge de philologie et d'histoire*, 82 (3), 733-762. https://doi.org/doi:10.3406/rbph.2004.4856

Lazard, G. (2009). Qu'est-ce qu'un sujet? *La linguistique*, 45 (1), 151-158.

Le Querler, N. (2001). La place du verbe modal pouvoir dans une typologie des modalités. In *Les verbes modaux* (pp. 17-32). Leyde : Brill.

Leeman-Bouix, D. (2002). *Grammaire du verbe français : des formes au sens : modes, aspects, temps, auxiliaires*. Paris : Nathan.

Marque-Pucheu, C. (2001). Valeurs de devoir dans les énoncés comportant selon N. In P. Dendale & J. Van der Auwera (Eds.), *Les verbes modaux* (pp. 85-101). Leyde : Brill.

Mayaffre, D., Pincemin, B., & Poudat, C. (2019). Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse de discours. *Langue française*, 3, 101-115.

Patard, A. (2015). Les constructions modales comparatives faire mieux de, valoir mieux et falloir mieux. *Syntaxe et sémantique*, 16 (1), 123-156. https://doi.org/10.3917/ss.016.0123

Reed, L. A. (2019). Further Implications of French Devoir and Falloir for Theories of Control and Modality. In D. L. Arteaga (Ed.), *Contributions of Romance Languages to Current Linguistic Theory* (pp. 65-89). Cham : Springer International Publishing. https://doi.org/10.1007/978-3-030-11006-2_4

Riegel, M., Pellat, J.-C., & Rioul, R. (2008). *Grammaire méthodique du français*. Paris : Presses Universitaires de France.

Rivière, C. (1984). Les équivalents anglais de «devoir» et «falloir». *Cahiers Charles V*, 6 (1), 7-27.

Rossari, C., Ricci, C., & Dolamic, L. (2018). Le conditionnel appliqué à devoir/dovere et son potentiel argumentatif. *Langue française*, 4, 105-120.

Salvadori, J., Barque, L., Haas, P., Huyghe, R., Lombard, A., Monney, M., Schwab, S., Tribout, D., & Wauquier, M. (2021). *The semantics of deverbal nouns in French: Annotation guide*. Fribourg : Université de Fribourg.

Sueur, J.-P. (1977). À propos des restrictions de sélection : les infinitifs devoir et pouvoir. *Lingvisticae investigationes*, 1 (2), 375-409.

Sueur, J.-P. (1979). Une analyse sémantique des verbes devoir et pouvoir. *Français (Le) Moderne Paris*, 47 (2), 97-120.

Tasmowski, L., & Dendale, P. (1994). Pouvoir E : un marqueur d'évidentialité. *Langue française*, 102, 41-55. http://www.jstor.org/stable/41559282

Valiukienė, V. (2016). *La multifonctionnalité et les effets de sens de nécessité des verbes modaux devoir et falloir et leurs équivalents en lituanien. L'étude basée sur le corpus français-lituanien/lituanien-français*. Thèse de doctorat. Vilniaus : Vilniaus universitetas.

Valiukienė, V. (2022). Du verbe plein à la composante pragmatique: le cas de falloir. *Verbum*, 13, 1-13. https://doi.org/10.15388/Verb.29

Varvara, R., Salvadori, J., & Huyghe, R. (2022, 2 septembre). Assessing affix polyfunctionality through BERT-derived representations. *20th International Morphology Meeting (IMM20)*, Budapest.

Vetters, C. (2004). Les verbes modaux pouvoir et devoir en français. *Revue belge de philologie et d'histoire*, 82 (3), 657-671. https://doi.org/doi:10.3406/rbph.2004.4851

Vetters, C. (2012). Modalité et évidentialité dans pouvoir et devoir : typologie et discussions. *Langue française*, 1, 31-47.

Vetters, C., & Barbet, C. (2006). Les emplois temporels des verbes modaux en français : le cas de devoir. *Cahiers de praxématique*, 47, 191-214.

Création et codage d'un corpus multimodal de repas familiaux

Christophe Parisse[1], Marion Blondel[2], Stéphanie Caët[3,2], Claire Danet[4,2],
Sophie de Pontonx[1] et Aliyah Morgenstern[5]
[1] Laboratoire MODYCO, Université Paris Nanterre
[2] Laboratoire SFL, CNRS-Paris8
[3] STL, CNRS et Université de Lille
[4] Dylis, Université de Normandie, Rouen
[5] Laboratoire PRISMES, Université Paris III Sorbonne Nouvelle
cparisse@parisnanterre.fr, marion.blondel@cnrs.fr, stephanie.caet@univ-lille.fr, claire.danet@gmail.com,
sdepontonx@parisnanterre.fr, aliyah.morgenstern@sorbonne-nouvelle.fr

Introduction

Le projet ANR DINLANG a pour objectif d'étudier les situations de repas familiaux dans des familles d'au moins quatre personnes utilisant au quotidien a) soit une langue majoritairement visuo-gestuelle, la LSF (Langue des Signes Française), b) soit une langue majoritairement audio-vocale, le français. Notre but est d'améliorer nos connaissances sur les liens entre formes et pratiques langagières multimodales, pratiques dinatoires, culture et développement du langage. Ce projet défend des choix théoriques forts qui portent sur les interactions langagières spontanées en milieu familial :

- Elles impliquent plusieurs partenaires dont les rôles conversationnels varient de manière dynamique.
- Elles engagent toutes les ressources sémiotiques à la disposition des participants, qu'elles soient visuo-gestuelles ou audio-vocales. Les gestes/signes, les bruits/paroles, les regards, les expressions faciales forment un tout et il y a une continuité entre toutes les productions plutôt qu'une division en sous-systèmes.
- Elles prennent sens dans la dynamique conversationnelle et actionnelle de chaque diner et dans l'histoire des interactions partagées entre les membres de la famille.
- Elles ne font sens qu'en situation, en relation avec le corps, l'environnement et les comportements des participants.

Ces principes sont à la croisée d'une série d'approches théoriques initiées ou présentées dans les travaux de Goldberg (2006), Linell (2009), Bottineau (2012), Mondada (2016), Morgenstern et al. (2021), Morgenstern (2022), complétées par une approche de la gestualité inspirée des travaux de Boutet (2018).

Nos recherches antérieures nous ont incités à construire un corpus multimodal, multi-perspectives, et un schéma d'annotation multi-pistes, permettant des codages et des analyses sur des dimensions très variées. Nos travaux (Morgenstern, 2014 ; Blondel et al., 2017 entre autres) sur les langues des signes nous ont confortés dans l'idée d'analyser le rôle du corps (mains, bras, buste, tête, regard) pour comprendre aussi bien le fonctionnement des langues gestuelles que des langues dites 'parlées'.

Notre corpus est basé sur des situations d'usage spontané du langage et de communication, dans le contexte d'une activité quotidienne, les repas familiaux français. Notre choix est dicté

par les raisons suivantes : il s'agit de situations d'une grande importance sociale et culturelle ; les interactions y sont multiples et non limitées à deux participants ; parents et enfants sont associés dans des situations partagées et de transmission de pratiques (langagières, dinatoires, etc.) ; ces moments d'interactions sont assez « ordinaires » pour ne pas être ressentis comme trop intimes pour être partagés dans le cadre d'un programme de recherche[95].

Le recueil de données

Le recueil s'effectue dans des familles comprenant les parents et deux ou trois enfants dont un au moins dans une tranche d'âge allant de 3 à 10 ans. Pour altérer le moins possible le déroulement habituel du diner, nous avons utilisé le dispositif suivant (Figure 1) :

- deux caméras de qualité et disposées de façon à permettre une prise de vue depuis la gauche de la scène du repas et une prise de vue depuis la droite de la scène ;
- une caméra 360° posée au centre de la table et offrant une vue de face de tous les participants ;
- un enregistreur sonore 360° posé sous la caméra 360°.

Un montage est effectué ensuite pour aligner temporellement tous les médias. Des extractions de la caméra 360° sont réalisées pour visualiser au mieux les visages et le haut du corps des participants. Les corpus constitués avec consentement éclairé des participantes seront rendus accessibles aux membres de la communauté scientifique à la fin du programme de recherche.

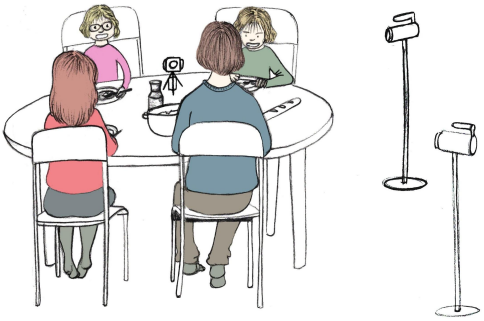


Fig. 1 : Plan d'installation du matériel de capture audiovisuelle

Le codage des données

Les questions de recherche qui sont posées dans notre projet orientent naturellement nos codages et nos analyses. Ainsi, nous voudrions notamment répondre aux questions suivantes :

- Y a-t-il des différences entre la pratique des activités dinatoires et la pratique des activités langagières des familles utilisant la LSF ou le français (notamment dans l'utilisation du regard, des mouvements des mains, des bras, du visage, de la bouche) ?
- En dépit des variations individuelles, trouve-t-on des tendances générales de développement des modes d'expression selon l'âge des participants ?

[95] Le projet a fait l'objet d'un examen par le Comité d'Éthique de la Recherche de la Sorbonne Nouvelle et les données ne sont analysées puis partagées et archivées qu'avec le consentement éclairé des participants avec qui nous partageons les données filmées dans leur famille.

- Lorsqu'ils sont utilisés en contexte, les modes d'expression ont-ils un impact majeur sur la manière dont les enfants construisent le sens et développent le langage ?
- Les familles qui communiquent en LSF ou en français emploient-elles les mêmes stratégies pour co-articuler les activités dinatoires et langagières ?

Afin de traiter ces questions, nous codons les actions et pratiques langagières, les destinataires des actions et productions langagières, les regards, le cadre participatif (qui participe à l'interaction), le thème traité, et ceci dans les deux modalités (audio-vocale et/ou visuo-gestuelle). Une contrainte est que la plupart des éléments analysés sont codés de manière indépendante au niveau temporel. Ainsi, la temporalité d'un geste, d'un regard, d'une parole, d'une action, même s'ils peuvent être considérés comme étant parfois interdépendants (au niveau intra-individuel et inter-individuel), sont différents dans le détail de leur production.

Nous avons utilisé un outil très courant dans l'analyse de la multimodalité, ELAN (2021), et créé un schéma d'annotation, ou *template* dans ce logiciel, ainsi qu'un manuel de codage permettant de coder tous les paramètres indiqués ci-dessus. Ce *template* et le manuel de codage seront fournis avec le corpus et des exemples de codages seront présentés lors des JLC.

L'analyse des données

Une des difficultés majeures rencontrées à la suite des codages réalisés est celle de l'analyse quantitative et multifactorielle des données (statistiques descriptives ou inférentielles). En effet, l'absence de correspondance temporelle entre les différents éléments codés dans les pistes d'ELAN ne permet pas de créer aisément une structure de type multivariée permettant de réaliser des analyses et des statistiques. Par exemple, pour savoir si un regard a un lien avec une action ou une production langagière, il faut d'abord mettre en relation les éléments potentiellement asynchrones des pistes « regard », « action », « pratique langagière[96] » avant de pouvoir étudier les caractéristiques qui les lient. On a donc besoin d'une mise en relation d'éléments répartis sur plusieurs pistes sur la base de leur relation temporelle, et sans codage de la relation de dépendance comme le logiciel ELAN permet d'en intégrer dans la structure d'annotation. Cette mise en relation peut se faire de deux manières que nous avons testées. La première méthode est d'utiliser la « recherche structurée » du logiciel ELAN, technique extrêmement puissante qui permet de nombreuses variations de contraintes dans les requêtes à appliquer sur les différentes pistes. Par exemple, les contraintes peuvent porter sur la temporalité (avant, après, pendant, etc.), mais aussi sur des éléments inclus dans les séquences à rechercher, sur la sélection des pistes à explorer, etc. La méthode permet l'export sous un format CSV destiné à un tableur ou à un logiciel de statistiques. Cependant, cette méthode gère mal les contraintes sur plus de deux pistes à la fois[97]. Une autre méthode plus compliquée à mettre en œuvre mais nettement plus puissante et rapide à l'usage est de construire un programme Python exploitant des bibliothèques permettant l'analyse de fichiers issus de ELAN. Cette méthode est plus puissante car le potentiel d'expression d'un langage de programmation est infini. Par contre, il faut réaliser un programme spécifique pour toute

[96] Nous appelons « pratiques langagières » en français, ce que nous avons libellé dans notre codage en anglais comme étant du « languaging » (Linell, 2009) et qui est catégorisé en modalités, du point de vue du codeur, auditif (lang aud : paroles, onomatopées, rires…) et visuel (lang vis : gestes, signes, expressions faciales). Le regard est codé sur une piste séparée.

[97] Techniquement, la limitation vient du fait qu'une combinaison de contraintes sur deux éléments n'est pas équivalente à des contraintes portant sur 3 ou 4 éléments à la fois.

interrogation, ou créer des programmes plus complexes avec des paramètres multiples, en évitant de reproduire ce qui existe déjà dans ELAN. Dans le cadre du programme DINLANG, nous avons choisi l'option des programmes spécifiques. Ils ont l'avantage par rapport à l'outil intégré dans ELAN d'être rapides et efficaces, mais l'inconvénient de ne faire qu'un seul type de calcul par outil. Ils sont disponibles sous forme de service web facilement utilisables sans connaissance informatique (voir https://ct3.ortolang.fr/toolselan/) ou sous forme de bibliothèque Python.

Conclusion

La création et l'exploitation de corpus multimodaux riches sont rendues possibles aujourd'hui grâce à l'amélioration des outils d'enregistrement et celle des outils informatiques. Il s'agit néanmoins d'un travail de longue haleine nécessitant des compétences spécifiques à développer collectivement. Pour cette raison, il nous semble possible et souhaitable que les techniques que nous avons utilisées pour l'analyse de nos données collectées durant des diners familiaux et dont nous ferons la démonstration lors de notre présentation soient partagées et utilisées de manière similaire dans d'autres projets portant sur la multimodalité des interactions langagières.

Références bibliographiques

Blondel, M., Boutet, D., Beaupoil-Hourdel, P., Morgenstern, A. (2017). La négation chez les enfants signeurs et non signeurs : des patrons gestuels communs, *LIA,* 8(1) : 141-171.

Bottineau, D. (2012) Parole, corporéité, individu et société : l'embodiment dans les linguistiques cognitives. *Texto !,* vol. XVII, n°1 and 2.

Boutet, D. (2018). *Pour une approche kinésiologique de la gestualité*. (Habilitation à diriger des recherches). Université de Rouen-Normandie. https://hal.archives-ouvertes.fr/tel-02357282

ELAN (Version 6.2) [Computer software]. (2021). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Goldberg, A. (2006). *Constructions at Work: The Nature of Generalization in Language*, Oxford, Oxford University Press.

Linell, P. (2009). *Rethinking Language, Mind and World dialogically: Interactional And Contextual Theories Of Human Sense-Making*. Charlotte, NC: Information Age Publishing.

Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Sociolinguistics*. Volume 20, issue 3: 336-366.

Morgenstern, A. (2014). Children's multimodal language development. In Christiane Fäcke (ED.). *Manual of language acquisition*. Berlin/Boston: De Gruyter.123-142. <halshs-01350596v1 >

Morgenstern, A. (2022). Children's multimodal language development from an interactional, usage-based, and cognitive perspective. *Wire's cognitive science.* https://doi.org/10.1002/wcs.1631

Morgenstern, A., Caët, S., Debras, C., Beaupoil-Hourdel, P, Le Mené, M. (2021). Children's socialization to multi-party interactive practices: Who talks to whom about what in family dinners. In Letizia Caronia (ed.) *Language and Social Interaction at Home and in School*. Amsterdam: John Benjamins, 46-85.

CORLI : Un corpus ouvert du français – ou comment travailler à rassembler les briques existantes ?

Christophe Parisse[1], Céline Poudat[2], Flora Badin[3], Christophe Benzitoun[4], Sascha Diwersy[5], Carole Etienne[6], Julie Glikman[7], Marie-Paule Jacques[8], Amalia Todirascu[9], Agnès Tutin[10].

[1] Laboratoire MODYCO, CNRS & Université Paris Nanterre

[2] Laboratoire BCL, Université Côte d'Azur

[3] Laboratoire LLL, CNRS & Université d'Orléans

[4] ATILF, Université de Lorraine

[5] Praxiling, Université Paul-Valéry - Montpellier 3,

[6] ICAR, CNRS

[7,9] Université de Strasbourg

[8,10] LIDILEM, Université Grenoble Alpes

cparisse@parisnanterre.fr[1], celine.poudat@univ-cotedazur.fr[2], flora.badin@univ-orleans.fr[3], christophe.benzitoun@univ-lorraine.fr[4], sascha.diwersy@univ-montp3.fr[5], carole.etienne@ens-lyon.fr[6], glikman@unistra.fr[7], marie-paule.jacques@univ-grenoble-alpes.fr[8], todiras@unistra.fr[9], agnes.tutin@univ-grenoble-alpes.fr[10]

Construire un corpus ouvert du français à partir de l'existant

L'idée de disposer d'un corpus représentatif et volumineux de la langue française et qui soit mis librement à la disposition des linguistes avec des outils permettant de s'en servir n'est pas nouvelle. Depuis plus d'une vingtaine d'années, avec l'accès facile à Internet et la diminution du coût des supports et outils informatiques, de nombreuses initiatives ont été lancées et ont permis de rendre disponible tout un ensemble de corpus de langue française, tant pour des données écrites que pour des données orales. Ainsi on pourra trouver dans la bibliographie une liste d'initiatives françaises de ce type.

Cette série de corpus est loin d'être exhaustive mais elle est représentative de la situation actuelle des corpus de français. Ces travaux ont permis de mettre à disposition de nombreux chercheurs des données et des outils. Ils ont tous une grande valeur car ils ont été produits par des spécialistes des sciences du langage et ils ont pour la plupart nécessité un investissement important, tant au niveau des heures de travail engagées qu'en termes financiers. Malheureusement, ces travaux et ces initiatives, quelle que soit leur grande valeur individuelle, reposent sur des standards parfois différents, ne sont pas accessibles avec les mêmes outils ou sur les mêmes sites. De plus, pour nombre d'entre eux, leur survie dans le temps n'est pas assurée et ils sont figés dans leur développement, du fait par exemple de la fin du projet qui a permis de les financer.

Idéalement, on souhaiterait que tous les corpus créés par des équipes de linguistes soient rassemblés en un seul lieu, utilisent des standards communs et soient accessibles de manière transparente avec des outils multiplateformes puissants, récents et maintenus. Si c'était le cas, la question de la création d'un corpus de « référence » se poserait moins car un grand corpus serait disponible pour la communauté. De plus, si ce corpus disposait de métadonnées, les chercheurs pourraient facilement contraster leurs propres données avec d'autres ensembles choisis, ou se constituer des sous-corpus d'étude selon leurs objectifs de recherche.

La tâche peut sembler complexe et ardue, mais nous pensons aujourd'hui être dans un contexte favorable pour mettre en œuvre un projet de grand corpus ouvert du français. Outre le fait que différentes initiatives de standardisation et d'agrégation de corpus existants ont vu le jour avec succès ces dernières années (e.g. CEFC-Orféo, CoMéRé, E-Calm) et que nous pouvons donc nous appuyer sur leurs expériences, les distances entre les corpus, les formats et les métadonnées, sont souvent plus courtes qu'on pourrait le craindre. Ceci s'explique par le fait que les formats aujourd'hui utilisés ne sont pas si nombreux que cela, la majorité des usages étant le texte brut, le format XML-TEI, ou un format tabulaire de type CONLL. Il en est de même pour les métadonnées. Elles sont certes très variées, mais les questions posées lors de leur création se rejoignent le plus souvent et des convergences, au moins pour les métadonnées les plus utiles, sont souvent possibles. Enfin, il existe maintenant de nombreux outils de recherche et d'analyse, qui de plus savent intégrer des corpus dans des formats variés texte, tableur, XML/TEI.

C'est dans ce contexte que le consortium CORLI, grâce au soutien de l'infrastructure Huma-Num, peut apporter des réponses et des moyens en vue de cet objectif. CORLI a montré depuis plusieurs années que les chercheurs ont très envie de collaborer et de partager leurs connaissances, leurs outils et leurs questionnements autour des corpus. CORLI a la chance de ne pas avoir des contraintes budgétaires similaires aux ANR. Au contraire, la fonction de CORLI (comme celle d'autres consortiums) est bien de produire ou d'aider à produire du matériel technique qui soit utilisé par tous et ceci sans objectif de compétition dans la recherche. CORLI peut donc, et le montre depuis plusieurs années, faire collaborer les chercheurs en sciences du langage autour d'un objectif commun, ce que nous nommons ici le « Corpus Ouvert du Français », ou « Open French Corpus » (OFC) pour une appellation internationale - souhaitée par notre sponsor Huma-Num et par nos institutions.

Premières réalisations : des liens et des briques complémentaires

Le travail sur le OFC a commencé à CORLI en 2022 et, nous souhaitons le continuer au moins jusqu'en 2025. Ce calendrier ne nous empêche pas de vouloir produire un résultat visible aussitôt que possible et donc bien avant 2025, même s'il sera amélioré par la suite. Notre cahier des charges très simplifié est le suivant :

- Utiliser les ressources existantes, corpus et outils, pour minimiser la création de données ou outils originaux ;
- Rassembler les corpus existants sur les sites de références et les projets autonomes ;
- Fournir des interfaces communes pour faire des recherches sur les corpus, y accéder, les télécharger ;

- Commencer le travail par les formats les plus simples (texte brut et métadonnées les plus fréquemment renseignées) de telle façon à rendre l'OFC disponible aussitôt que possible et l'enrichir par la suite au fur et à mesure.

Le travail de CORLI a débuté en 2022 par une réflexion sur les métadonnées à choisir et privilégier pour les corpus, notamment de langue écrite (une réflexion sur les métadonnées de l'oral existe déjà, fruit du travail de CORLI dans les années précédentes). Ce travail est réalisé en collaboration avec le consortium ARIANE (anciennement CAHIER).

Une deuxième partie du travail à réaliser comprend une harmonisation des formats de données qui permette leur exploitation par des outils informatiques. CORLI a mené en 2022 une étude sur les formats des données déposées dans ORTOLANG, et on s'aperçoit que, fort heureusement, le choix des formats est moins disparate qu'on ne pourrait le craindre. De nombreuses données sont au format texte ou TEI, ou facilement convertibles dans ces formats (en grande partie grâce au travail passé de CORLI ou du consortium TEI). Également, l'agrégation de corpus issus de sources multiples pour générer un corpus d'étude en TEI, réutilisable dans des outils comme TXM a déjà été démontrée (Parisse et al. 2021).

La combinaison des métadonnées et de la TEI pour enrichir les données de corpus est déjà possible, comme démontrée et outillée dans le travail de Badin et al. (2021). Les outils utilisés existent déjà et sont disponibles dans l'outil CORLI TEICORPO (Parisse, Etienne & Liégeois, 2020).

L'interrogation de ces données rassemblées en un seul corpus peut se faire avec de nombreux outils existants. CORLI, avec l'appui d'Huma-Num et des dépôts COCOON et ORTOLANG, a déjà incité à la mise en place dans COCOON et dans ORTOLANG du *ContentSearch* de CLARIN qui permet une interrogation des textes bruts déposés dans ces dépôts. Il manque pour la généralisation de cet outil des conversions systématiques des formats, qui seront possible grâce aux outils évoqués dans le paragraphe précédent. Par ailleurs, un travail en association entre CORLI et ORTOLANG doit débuter cette année sous la direction de l'ATILF pour utiliser le moteur de recherche de Frantext (Allegro) sur les données d'ORTOLANG, après conversion et filtrage des documents conformément au travail réalisé par CORLI.

Le travail planifié consiste clairement à faire des liens entre des briques existantes et donc à créer le moins de code ou outil informatique nouveau. Malgré tout, une certaine complexité existe dans notre projet. C'est pourquoi notre désir est aussi de procéder par étape en commençant par fournir des textes au format brut UNICODE, avec des métadonnées minimales (propriétaire, droit, année). La deuxième étape sera d'injecter des données issues, non seulement d'ORTOLANG, mais aussi de tout type de source. Format et métadonnées seront ensuite enrichis, puis nous pourrons aussi ajouter les informations grammaticales ou sémantiques (voir par exemple https://corliapi.ortolang.fr/), et tout autre type d'annotation de corpus pour bénéficier des résultats des autres projets du consortium CORLI (voir la liste des projets sur notre site : https://corli.huma-num.fr/projets-corli-2022-2025/).

Nous sommes conscients que du projet à la réalisation un certain nombre d'obstacles peuvent se présenter, mais nous sommes confiants dans la capacité d'adaptation de notre consortium, en particulier grâce à la grande variété de compétences humaines de son comité de pilotage et de son conseil scientifique. Nous pensons également que l'aspect collectif et consensuel de de la création de ce corpus ouvert est le meilleur gage de son succès futur.

Références bibliographiques

Badin, F., Liégeois, L., Thiberge, G., & Parisse, C. (2021). Vers un outillage informatique optimisé pour corpus langagiers oraux en vue d'une exploitation textométrique: Le cas des interrogatives partielles dans ESLO. Corpus, 22, Article 22. https://doi.org/10.4000/corpus.5752

Parisse C., Benzitoun C., Étienne C. et Liégeois L. (2021). Agrégation automatisée de corpus de français parlé In : Des corpus numériques à l'analyse linguistique en langues de spécialité [en ligne]. Grenoble : UGA Éditions, 2021. Disponible sur Internet : <http://books.openedition.org/ugaeditions/24220>. ISBN : 9782377473021. DOI : https://doi.org/10.4000/books.ugaeditions.24220.

Parisse, C., Etienne, C., & Liégeois, L. (2020). TEICORPO: A Conversion Tool for Spoken Language Transcription with a Pivot File in TEI. Journal of the Text Encoding Initiative, Issue 13, Article Issue 13. https://doi.org/10.4000/jtei.3464

Exemples de corpus accessibles

Corpus écrits

Scientext https://scientext.hypotheses.org/corpus

Scienquest https://corpora.aiakide.net/

Archives parlementaires https://archives-parlementaires.persee.fr/

Consortium CAHIER https://cahier.hypotheses.org/

E-CALM https://www.ortolang.fr/market/corpora/e-calm

Corpus 14 https://www.univ-montp3.fr/corpus14/

Democrat https://hdl.handle.net/11403/democrat

Corpus oraux

CFPP2000 http://cfpp2000.univ-paris3.fr/search.html

ESLO http://eslo.huma-num.fr/index.php/pagecorpus/pageaccescorpus

CHILDES https://talkbank.org/DB/

CT3-ORTOLANG https://ct3xq.ortolang.fr/ct3xq/interro

PFC https://public.projet-pfc.net/transcription/

Corpus multimodaux

CEFC-Orféo https://orfeo.ortolang.fr/, http://orfeo.grew.fr/

CoMeRe http://hdl.handle.net/11403/comere

Autres types d'initiatives, corpus semi-ouverts

Frantext https://www.frantext.fr/repository/frantext-demo

Les bases d'Hyperbase (E. Brunet) http://ancilla.unice.fr/pages/bases/

« De l'exploitation d'un corpus numérique à l'enseignement d'une notion théorique en licence professionnelle »

Eugénie Pereira Couttolenc
Laboratoire Éducation Discours et Apprentissages (EDA), Université Paris Cité
eugeniepereira@hotmail.fr

Mots-clés : présentation de soi, analyse du discours, corpus numérique, approches linguistiques, application, licence professionnelle.

Résumé de la communication

En tant qu'exemple de retour d'expérience et dans une perspective didactique, cette communication prétend illustrer l'appréhension d'une notion théorique en sciences du langage au moyen de la constitution d'un corpus de données natives du web et dans le cadre d'un enseignement dispensé auprès d'un public d'étudiant.e.s de 3ème année de licence professionnelle préparant aux métiers de la rédaction et de la communication multimédia.

Le contexte

Tout d'abord, pour contextualiser notre cas, il est important d'indiquer qu'au cours de notre mission d'enseignement réalisée auprès de l'Université Paris-Est Créteil (UPEC) et dans une perspective d'application professionnelle, il nous a été demandé de sensibiliser des étudiant.e.s de troisième année de licence sur le thème du « savoir argumenter par le texte ». Face à un public non formé aux sciences du langage – puisque composé de jeunes professionnel.le.s provenant de filières techniques dans les secteurs de la vente, de la communication ou de l'audiovisuelle – il s'est révélé nécessaire de porter un soin particulier au choix des supports visant à transmettre le contenu théorique la matière. Ainsi, si les affiches publicitaires ou les textes à visée promotionnelle sont effectivement des supports pertinents pour appréhender, par exemple, les différents types d'arguments répertoriés en rhétorique, leur dimension majoritairement plurisémiotique se révèle plus adaptée à un enseignement ayant pour objectif de « savoir argumenter par l'image » (un enseignement qui, de fait, est dispensé au second semestre au sein de ce même parcours). Par ailleurs, rejoignant les postures de la linguistique de corpus (Teubert, 2009 : 185-211) et les propos de Surcouf et de Blanche- Benveniste sur l'intérêt de proposer des ensembles composés de documents authentiques et contemporains pour transmettre des savoirs linguistiques (Blanche-Benveniste, 2003 : 317 ; Surcouf, 2021 : 108), j'ai choisi de travailler la notion de *présentation de soi discursive* (Amossy, 1999 et 2010) à partir d'un ensemble de productions introduisant la personnalité et le parcours d'animateur.trice.s de blogs voyage actif.ve.s sur les réseaux sociaux francophones.

Le sujet de l'enseignement : la présentation de soi

De fait, on a établi, lors d'un précédent travail de recherche mené en analyse du discours, qu'au moment de la prise de parole des blogueur.euse.s, ceux-ci ne bénéficient ni de la crédibilité ni de l'autorité nécessaires pour prodiguer des conseils et des recommandations

touristiques (Pereira Couttolenc, 2022 : 55). En tant qu'anonymes, il leur est donc essentiel de « se construire une position de légitimité » (Charaudeau, 2002 : 340). Cette démarche singulière, visant à convaincre et à persuader un auditoire, se produit en divers endroits constitutifs de la structure préétablie des blogs et plus spécifiquement au cœur des sections « À propos », « Qui-suis-je ? » ou encore « Qui se cache derrière ce blog ? ». Or, les interventions des auteur.e.s de blogs voyage portent en elles des indices sur le style, sur les compétences langagières et encyclopédiques et sur les croyances de leurs locuteur.trice.s et de leur public (Amossy, 1999 : 9-11 ; Kerbrat-Orecchioni, 1980 : 20). Ainsi, la personnalité que le sujet parlant dévoile de lui-même, l'ethos qu'il manifeste, les représentations qu'il se fait de son objet de discours, de ses interlocuteurs, des attentes de son public se dénotent de la manière dont celui-ci mobilise la langue et ses possibilités (Culioli, 1999 : 92 ; Kerbrat-Orecchioni, 1980 : 20 ; Maingueneau, 2013 : 88). L'allocutaire – ici l'internaute – peut alors, consciemment ou non, reconstituer le « corps énonçant » (Maingueneau, 1999 : 76) de l'instance discursive du locuteur (Amossy, 1999 : 18). Dans le cadre de cette étude, on a adopté le concept de *présentation de soi en ligne* (Pereira-Couttolenc, 2022 : 57) qui réunit l'ensemble des éléments techniques, langagiers et sémiotiques concourant à la production de l'image de soi au sein d'un objet natif du numérique.

Application dans un parcours de licence professionnelle

En accord avec les précédentes affirmations théoriques et méthodologiques, nous avons composé en 2022 pour les étudiant.e.s de la troisième année de la licence parcours Rédaction Professionnelle et Communication Multimédia un corpus de travail composé de billets de présentation d'influenceur.se.s voyage à partir duquel il était possible d'observer, d'identifier et de décrire les marques sémiotiques et langagières témoignant d'une démarche argumentative. Guidés dans l'exploitation des données, puis confrontés à d'autres corpus réunissant des guides pratiques et des brochures touristiques, les étudiant.e.s ont pu reconstituer, à partir de ces différents marqueurs, la construction en discours de la figure de l'expert du domaine viatique. Par ailleurs, afin d'évaluer les acquis de cet enseignement, les apprenants ont été invités à jouer, à leur tour, le rôle d'influenceur.euse dans le domaine de loisirs de leur choix. Pour ce faire, la première partie de la consigne indiquait la nécessité de sélectionner un secteur d'activité et d'analyser les productions de trois influenceur.euse.s actif.ve.s sur le segment retenu. Puis, les étudiant.e.s devaient spécifier leurs intentions en termes de positionnement de leur site, de l'image de soi qu'ils aspiraient transmettre et du public qu'ils souhaitaient atteindre avant de proposer un texte introductif visant à convaincre et à séduire de potentiels lecteur.trice.s.

Nous avons fait le choix de recourir à la notion de présentation de soi en discours pour amener notre public de jeunes adultes à se sensibiliser, d'une part, à l'importance de l'image discursive du.de la locuteur.trice transmise au sein d'un texte argumentatif et, d'autre part, aux stratégies sémiotiques et langagières qu'il est possible de déployer pour se forger une crédibilité auprès d'un large public d'internautes. Notre exposé sera illustré d'exemples issus du corpus de travail ayant servi à l'élaboration du contenu de l'enseignement. Il sera également complété d'extraits de productions des étudiant.e.s qui semblent, vraisemblablement, avoir été réceptif.ve.s tant à la démarche qu'à la finalité de l'exercice.

Bibliographie

Amossy, R. ([2010] 2015). *La présentation de soi – Ethos et identité verbale*, Paris : P.U.F.
Amossy, R. (1999). *Images de soi dans le discours – la construction de l'ethos*. Lausanne : Delachaux et Niestlé.

Blanche-Benveniste, C. (2003). « La langue parlée » *In* Yaguello, M. (dir.), *Le Grand Livre de la Langue française*. Paris : Seuil. 317-344.

Charaudeau, P. (2002). « Légitimation (stratégie de-) », *In* Charaudeau, P., Maingueneau, D., *Dictionnaire d'analyse du discours*. Paris : Seuil. 339-340.

Culioli, A. (1999). *Pour une linguistique de l'énonciation – domaine notionnel*. Paris : Ophrys. Kerbrat-Orecchioni, C. (1980). *L'Énonciation. De la subjectivité dans le langage*. Paris : Armand Colin.

Maingueneau, D. ([2013] 2021). *Analyser les textes de communication* (4e édition). Paris : Armand Colin.

Maingueneau, D. (1999). « Ethos, scénographie, incorporation », *in* R. Amossy, *Images de soi dans le discours – La construction de l'ethos*. Lausanne : Delachaux et Niestlé. 75-100.

Pereira Couttolenc. E. (2022). « La présentation de soi des auteur.e.s de blogs voyage : une identité discursive du voyageur amateur ? », thèse de doctorat en sciences du langage, dirigée par von Münchow, Patricia, Université Paris Cité, soutenue à Paris le 2 décembre 2022.

Surcouf, C. (2021). « Le français oral quotidien, un objectif spécifique en FLE ? Retour sur les défis de la création d'un corpus de français parlé annoté à visée pédagogique », *In* Frérot, C. et Pecman, M. (dir.), *Des corpus numériques à l'analyse linguistique en langues de spécialité*, Grenoble : UGA Éditions. 107-133.

Wolfgang, T. (2009). « La linguistique de corpus : une alternative [version abrégée] », *Semen* [En ligne], 27 | 2009, mis en ligne le 01 avril 2009, consulté le 16 février 2023. URL : http://journals.openedition.org/semen/8914 ; DOI : https://doi.org/10.4000/semen.8914

(Avoir) le QI de... - la syntaxe, la sémantique et la pragmatique d'une collocation intensifieuse non standard en français contemporain

Ewa Pilecka [1] et Tomasz Januchta [1]
[1] Université de Varsovie
e.pilecka@uw.edu.pl,t.januchta@student.uw.edu.pl

Introduction

Dans le cadre de notre projet de recherche qui a pour but l'élaboration d'un dictionnaire électronique des moyens d'intensification, nous nous intéressons entre autres à des intensifieurs « non-standard », qu'on ne trouve pas dans les dictionnaires traditionnels, mais dont la fréquence dans les corpus est suffisamment signifiante pour en faire de bons candidats à figurer dans un dictionnaire basé sur l'usage contemporain de la langue « dans tous ses états ».

Notre intervention portera sur la collocation « (avoir) le QI de Dét N » (cf. Romero 2017, fiche 6 : « avoit le QI d'un acarien »), que l'on peut la mettre en parallèle avec la comparaison intensifiante « être bête/con comme (Dét) N », qui à son tour est un cas particulier d'une structure suscitant depuis longtemps l'intérêt des linguistes (cf. Dauzat 1945, Buvet & Gross 1995, Gross 1996, Schapira 2000, Pierrard & Léard 2004, Leroy 2004, Izert 2002, Romero 2016) et des lexicographes (Cazelles 1996). Cependant, outre la ressemblance qui consiste à à intensifier la propriété 'bêtise humaine', les deux constructions diffèrent tant au niveau syntaxique que sémantique et pragmatique.

Ces trois niveaux de description feront l'objet d'un examen approfondi, qui nous amènera entre autres à :

- identifier les variations syntaxiques de la structure de base (« Dét QI de Dét N), qui portent en particulier sur :
    - la présence/absence d'un verbe support (avoir ou ses synonymes)
    - les formes des déterminants
    - les expansions du nom-parangon (syntagme adjectival, syntagme prépositionnel, proposition subordonnée, apposition...)
- dégager et classer les parangons de bêtise susceptibles d'apparaître comme Dét N (ex. volaille, animaux aquatiques, artefacts ménagers etc), qui diffèrent par ailleurs du paradigme présent dans la comparaison « bête comme... » ;

- décrire l'effet de surintensification et/ou d'expressivité réalisé à travers l'enrichissement syntaxico-sémantique de la structure de base (ex. (avoir) le QI d'une huitre / le QI d'une huitre sans perle / le QI d'une huître anémique mononeuronale / le QI d'une huitre en échec scolaire sous acide / d'une huitre albinos souffrant de cataracte et des pieds plats / le QI d'une fesse d'huitre / le QI d'un bulot, huître et autre crustacés etc) ;

- en étudier les valeurs pragmatiques (en particulier, celle de l'insulte, ainsi que de diverses stratégies de la prise en charge de celle-ci par l'énonciateur, qui se manifestent notamment dans le co-texte).

La recherche a été réalisée à partir du corpus frTenTen20 (composé de textes francophones collectés sur Internet entre 2019 et 2020 et comportant plus de 15 milliards de mots étiquetés en POS), accessible sur la plateforme SketchEngine. La requête en CQL (Corpus Query Language), avec la formule de départ [lemma="avoir"] [tag="D.*"] [word="QI"] [lemma="de"] [tag="D.*"] ? [tag="N.*"] a retourné 585 résultats, et celle où la présence du verbe n'est pas obligatoire, à savoir, [tag="D.*"] [word="QI"] [lemma="de"] [tag="D.*"] ? [tag="N.*"], a donné 1994 résultats (dont une partie cependant a dû être éliminé - soit manuellement, soit en modifiant la formule de requête - car ils constituaient des « bruits » du point de vue de l'objectif de notre recherche).

Résultats

Nous constatons une grande productivité syntaxique de la structure examinée, la liberté d'ajouter de nouveaux éléments grammaticaux à la version de base, ainsi que de les modifier et de les remplacer. La diversité syntaxique s'exprime ici non seulement dans la richesse des éléments potentiels qui peuvent être ajoutés, substitués et combinés de multiples façons, mais aussi parfois dans la longueur des structures. Le patron, bien qu'il ne soit pas rigide, ne se comporte pas de manière totalement libre, mais produit des faisceaux de ses actualisations, en particulier dans les combinaisons de noms et d'adjectifs spécifiques.

Nous notons également une gradation nuancée de l'intensification et de l'augmentation de l'expressivité, allant de pair avec la complication des structures syntaxiques. Le phénomène d'intensification, voire de surintensification, est en partie non corrélé avec les changements d'expressivité. Lorsque l'expressivité semble parfois continuer à augmenter avec la complexité syntactico-sémantique, ce n'est pas nécessairement le cas pour le degré d'intensification.

La stupidité est souvent représentée par des images d'un récipient fait de matière dure, vide, ce qui revient à se référer de manière anthropocentrique à une tête dépourvue de cerveau. A côté de la dureté et de la vacuité, ce sont les images de matières molles, animées ou non, qui dominent. Nous pouvons supposer que les traits opposés, c'est-à-dire la fermeté, la rapidité, la précision et la solidité, seraient des facteurs présents dans l'analyse de l'intelligence élevée.

Du point de vue pragmatique, la structure étudiée est le plus souvent utilisée soit pour offenser, soit à de fins d'autocritique. Le niveau d'imagerie et d'intensité contenu dans ces discours est très variable, et le caractère offensant procède des structures motivées rhétoriquement, créatives, multifacétiques, développées et même très élaborées.

Vu ce qui précède, nous considérons que l'inclusion de la formule avoir le QI de N dans les dictionnaires électroniques serait justifie, car elle refléterait le processus de lexicalisation en

cours, sinon achevé. Il serait également intéressant de procéder à des études contrastives avec d'autres langues où la même formule se rencontre (p.ex. l'anglais ou le polonais) afin de dégager les paradigmes de parangons, ce qui pourrait nous renseigner sur leur contexte culturel d'aujourd'hui.

Références bibliographiques

Bordas, É., 2022, La notion d'expressivité. Présentation, Langages, 228/4, 7-24.

Buvet P.-A. & Gross G., 1995, Comparaison et expression du haut degré dans le groupe nominal, Faits de langues, 5, 83-88.

Cazelles, N., 1996, Les comparaisons du français, Paris, Belin. Dauzat, A., 1945, L'expression de l'intensité par la comparaison, Le français moderne, 13/3-4, 169-186.

Fuchs, C., 2014, La comparaison et son expression en français, Paris-Gap, Ophrys.

Gross, G., 1996, Les expressions figées en français. Noms composés et autres locutions, Paris, Ophrys.

Izert, M., Pilecka, E., 2021, Comment « surintensifier » les expressions d'intensité ? L'exemple des collocations Adj /V comme SN et Adj/ N à faire V INF, Estudios Románicos, 30, 59-78.

Izert, M. (2002) Les expressions Adj comme SN et l'intensification de la propriété, thèse de doctorat, Université de Varsovie.

Leroy, S., 2004, Sale comme un peigne et méchant comme une teigne. Quelques remarques sur les comparaisons à parangon, Travaux linguistiques du cerlico, 17, 255-267.

Leroy, S., 2007, Les comparaisons comme SN exprimant le plus haut degré, Travaux de linguistique, 54, 69-82.

Mel'cuk, I., 2003, Collocations dans le dictionnaire, in : T.Szende (éd.) Les écarts culturels dans les dictionnaires bilingues, Paris, Honoré Champion, 19-64.

Pierrard, M. , Léard, J.-M., 2004, Comme : comparaison et haut degré, Travaux linguistiques du Cerlico, 17, 269-287.

Romero, C. , 2007, Pour une définition générale de l'intensité dans le langage, Travaux de linguistique, 54, 57-68.

Romero, C. , 2017, L'intensité et son expression en français, Paris, Ophrys. Romero, C.,

2015, A quoi compare-t-on pour intensifier ? in : K. Wróblewska-Pawlak, A.Kieliszczyk (éds) L'intensification et ses différents aspects, Warszawa, WUW, 133-152.

Schapira, C., 2000, Du prototype au stéréotype et inversement : le cliché comme + SN, Cahiers de lexicologie, 76, 27-40.

Tutin, A., Grossmann, F., 2002, Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif, Revue française de linguistique appliquée, 7, 7-25.

Un modèle pour décrire et annoter les discours autres

Céline Poudat[1], Marie Chandelier[1] et Gabriella de Luca [2]
[1] Laboratoire BCL, Université Côte d'Azur, France
[2] Laboratoire CEDITEC, Université Paris-Est Créteil, France
celine.poudat@univ-cotedazur.fr, marie.chandelier@univ-cotedazur.fr, gabriella.de-luca-silva-moreira@u-pec.fr

Introduction

Le développement des technologies et des réseaux sociaux s'est accompagné de l'augmentation de la production et de la circulation de contenus textuels. Les phénomènes de reprise, qui se sont alors multipliés, incarnent une composante majeure dans la construction formelle et sémantique des discours. Dans cette perspective, notre objectif est double :
(i) nous présentons un nouveau modèle d'annotation du discours rapporté construit sur le modèle théorique de Jacqueline Authier-Revuz (2020). Ce modèle de *Représentation du Discours Autre* nous semble pertinent pour mettre en évidence des phénomènes de reprise particulièrement fins tout en proposant de nouvelles perspectives de comparaison des genres discursifs ;
(ii) nous évaluons l'opérabilité et l'intérêt descriptif du modèle en annotant deux corpus de formes et de genres différents : un corpus d'articles de presse que nous comparons à un corpus de communication médiée par les réseaux, *i.e.* un corpus de discussions éditoriales écrites de Wikipédia. Ces deux corpus sont thématiquement articulés à deux grandes questions soci(ét)ales dont on interroge le traitement et la restitution : le phénomène des *gilets jaunes* pour les articles de presse et la crise climatique pour les discussions Wikipédia, qui sont représentatives des questionnements que pose l'écriture des articles encyclopédiques rédigés sur ces thématiques. S'il nous a semblé pertinent de tester le modèle sur des genres différents, ce qui est le cas ici puisque les deux corpus ne relèvent ni du même genre ni du même type (textuel *vs* conversationnel), nous nous sommes aussi intéressées aux propriétés partagées par ces deux genres : les articles de presse et les articles encyclopédiques sont en effet régis par un contrat de communication posant des enjeux similaires de véridicité et de crédibilité (Charaudeau 2005), ce qui rend cruciale la question des sources de l'information produite. Ces enjeux se retrouvent explicités et questionnés dans les interactions Wikipédia qui incarnent en quelque sorte des coulisses rédactionnelles auxquels on aurait difficilement accès pour les textes de presse, même coécrits

Le modèle d'annotation

Bien que le discours rapporté ait généré une littérature abondante (*e.g.* Rosier 2008, Holt in D'hont *et al.*, 2009; Coulmas 2011), la plupart des modèles exploités ne nous semblent pas permettre de saisir l'hétérogénéité énonciative des discours de manière aussi fine que le modèle d'Authier-Revuz (2020), soit parce qu'ils se limitent aux trois modes de discours rapportés encore usuellement enseignés dans le secondaire (discours direct, discours indirect, discours indirect libre), soit parce qu'ils se concentrent plutôt sur la source effective du dire en distinguant les paroles de L et l. La question des îlots textuels par exemple, pourtant très

répandue dans de nombreux genres médiatiques, est rarement prise en compte de manière pertinente:

Séisme en Turquie et en Syrie : l'OMS déplore le ***« pire désastre naturel en un siècle »*** *en Europe (Le Monde)*

Le modèle d'annotation que nous avons développé pallie cette absence en distinguant cinq modes de discours autre mis au jour par Authier-Revuz (2020) : le discours direct, le discours indirect et le discours indirect libre, auxquels s'adjoignent deux grands types de modalisation : la modalisation autonymique d'emprunt (*Julien Bayou estime que son « féminisme » est celui de « l'égalité entre les femmes et des hommes », qui évite les « dérives » et « excès » - Le Figaro*), et la modalisation en assertion seconde (*Selon la présidente de l'Assemblée nationale, Yaël Braun-Pivet, soixante-douze heures de débats ont été programmées - Le Monde*).

Deux autres paramètres ont été intégrés: (i) le caractère marqué ou non marqué de la forme, en fonction de la présence d'un marqueur de surface (morphologique ou syntaxique). Chaque mode de discours autre a des formes marquées et non marquées, à l'exception du discours indirect libre qui est fondamentalement non marqué ; et (ii) le degré de certitude des annotateurs pour les formes non marquées, qui sont bien sûr plus interprétatives. Une échelle de trois valeurs a ainsi été établie, selon que le degré de confiance était élevé, ambigu (possibilité de trancher mais reconnaissance d'une ambiguïté) ou indécidable (incapacité de trancher).

Les formes de discours autre ont été balisées à l'aide de la plateforme d'annotation *Inception* (TU Darmstadt). Intégrées dans le corpus, elles sont exprimées par des balises XML. Les segments sont annotés à l'aide d'une balise <rda> et leurs caractéristiques sont exprimées par les attributs suivants :

```
<rda mode="DD"|"DI"|"DIL"|"MAS"|"MAE"|"indéterminé"
certitude="élevée"|"ambigu"|"indécidable" forme="marquée"|"non marquée">...</rda>
```

: balises XML pour l'annotation

L'ensemble des deux corpus ont été annotés en suivant ces catégories par une annotatrice et systématiquement vérifiées par chacun de ses responsables, en vue d'obtenir un consensus et de stabiliser les annotations. Ce processus de recherche de consensus a été important pour définir l'annotation des cas les plus difficiles, comme le discours narrativisé et les hypothétiques pour les discussions Wikipédia ; ou le discours indirect libre pour les articles de presse. Dans les cas des discours narrativisé et des hypothétiques, les extraits ont été annotés comme du discours indirect marqué, en raison de la présence des verbes de parole :

[1] <rda mode="DI" certitude="élevée" forme="marquée">Mais Mr Lebrouillard/Tibo a parlé, c'est du vandalisme</rda> (discussion Wikipédia - Réchauffement climatique)

[2] <rda mode="DI" certitude=" élevée" forme="marquée">celui qui dit qu'il y a consensus scientifique autour de la thèse du réchauffement climatique d'origine anthropique, au mieux, se trompe, au pire, ment</rda> (discussion Wikipédia - Réchauffement climatique)

Le discours indirect libre étant invariablement interprétatif, la révision a été une étape essentielle pour décider l'annotation de ces formes :

> [3] <rda mode="DIL" certitude="élevé" forme="non marquée" >Comment éviter la voiture quand les interconnexions entre villes de banlieue sont encore quasi inexistantes? Quand la ligne de RER qui dessert sa ville est saturée?</rda> <rda mode="DD" certitude="élevé" forme="marquée">« Je prenais les transports en commun tous les jours quand je travaillais à la banque. Mais maintenant que j'ai créé mon entreprise, ils ne me permettent pas d'assurer mes rendez-vous quotidiens. »</rda> Priscilla Ludosky a lancé une boutique en ligne de cosmétiques bio.

Les corpus de travail

L'élaboration d'un modèle pour l'identification des séquences de discours représenté dans une variété de genres textuels présente comme nous l'avons vu un double intérêt. Notre contribution a pour objectif d'illustrer les apports de ce modèle (i) pour la description des caractéristiques des genres et (ii) pour la documentation du rôle joué par le discours autre dans la construction du point de vue énonciatif. Nous présenterons une analyse comparative de corpus relevant de deux genres pour lesquels les discours représentés occupent une place centrale : la presse écrite et les discussions Wikipédia.

Articles de presse

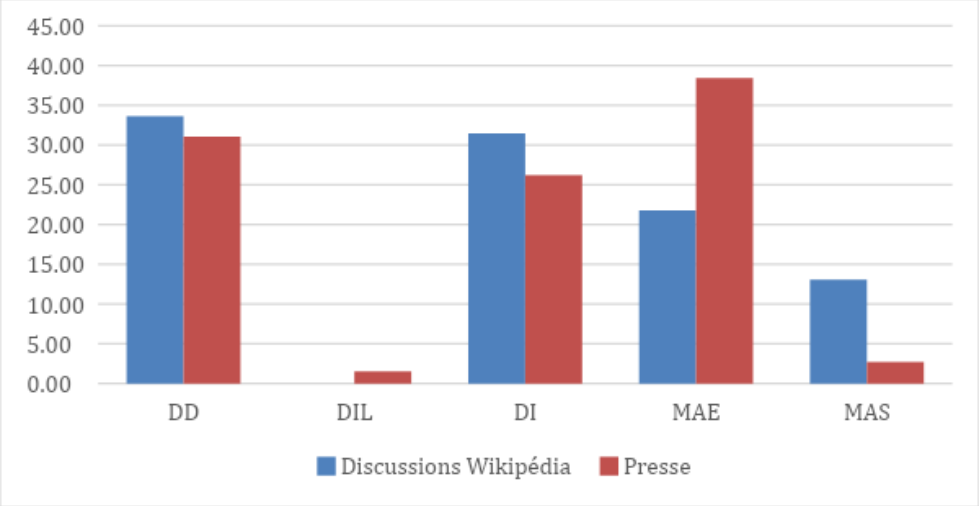
Le corpus de presse écrite est constitué d'articles consacrés au mouvement de contestation des Gilets jaunes provoqué par la hausse du prix du carburant (TICPE) en novembre 2018. Ce mouvement, qui s'est traduit par une forte mobilisation locale et nationale, a été abondamment relayé par la presse généraliste (Boyer et al. 2020). Nous avons annoté les 15 articles parus en novembre 2018 - premier mois de la contestation - dans le quotidien *Le Monde* (15 articles, 16,857 mots, 515 segments annotés).

Discussions Wikipédia

Le corpus Wikipédia a été établi en sélectionnant les discussions éditoriales associées à un ensemble d'articles emblématiques de la question climatique : *Réchauffement climatique, Effet de serre, Gaz à effet de serre* et *GIEC*. Ce corpus est intéressant à plus d'un titre : il permet d'observer comment s'instancie une controverse sociétale dans des échanges individuels dans le cadre très particulier des discussions Wikipédia (Poudat et al. 2017, Poudat & Ho-Dac 2019), régies par des règles de « neutralité de point de vue » (sic) et de sourçage très spécifiques. La question de la source est en effet cruciale dans l'encyclopédie et il nous semble très intéressant d'observer comment les sources sont introduites et comment les acteurs du débat climatique sont représentés. Le corpus agrège 208 fils de discussion regroupant 879 messages postés, ce qui représente 85798 mots.

Premiers résultats et perspectives

Le corpus de presse, cinq fois moins volumineux que le corpus de discussions Wikipédia, contient deux fois plus de marqueurs de discours rapporté, ce qui n'est pas surprenant si l'on s'en tient à la littérature : Alrahabi & Desclés estimaient déjà en 2008 que le discours rapporté pouvait constituer jusqu'à 90% des phrases d'un article. Dans les discussions Wikipédia, les échanges autour des sources sont fréquents mais évidemment moins omniprésents.



: Répartition des modes de discours autres dans les deux genres

Nous présenterons le détail des résultats et des comparaisons que nous avons réalisées mais de manière notable, on peut dire que les deux genres se caractérisent par une forte hétérogénéité énonciative montrée. Les enjeux de véridicité et de crédibilité des articles produits que nous avons déjà mentionnés conduisent les deux genres à privilégier les formes marquées du discours rapporté (Rosier 2002) ; aussi représentent-elles 97% des formes relevées dans le corpus de presse et 99,3% de celles relevées dans les discussions Wikipédia. Dans les deux cas, l'information relayée est explicitement attribuée à une source externe. De manière intéressante, la modalisation autonymique d'emprunt est davantage plébiscitée dans les articles de presse, attachés à la construction d'un *effet de réel* tandis que les Wikipédiens reformulent plus volontiers le contenu des sources citées, ce qui explique les plus grandes proportions observées de modalisation en assertion seconde et de discours indirect dans une moindre mesure (*v.* Figure 2).

Références bibliographiques

Alrahabi, M. & Desclés, J. (2008). Automatic Annotation of Direct Reported Speech in Arabic and French, According to a Semantic Map of Enunciative Modalities. In Nordström B., Ranta A. (eds) *Advances in Natural Language Processing*. GoTAL 2008. Lecture Notes in Computer Science, vol 5221. Berlin: Springer.

Authier-Revuz, J. (1984). Hétérogénéité(s) énonciative(s). *Langages*, 19(73), 98-111.

Authier-Revuz, J. (2020). La Représentation du Discours Autre : principes pour une description. Berlin : De Gruyter. coll. Linguistique française.

Boyer, P. C., Delemotte, T., Gauthier, G., Rollet, V., & Schmutz, B. (2020). Les déterminants de la mobilisation des Gilets jaunes. *Revue Économique*, 71(1), 109-138.

Charaudeau, P., (2005). *Les médias et l'information. L'impossible transparence du discours*, Louvain-la-Neuve, De Boeck-Ina.

Coulmas, F. (2011). Reported speech : Some general issues. In Reported speech : Some general issues. De Gruyter Mouton, pp. 1-28.

D'hondt, S., Östman, J.-O., & Verschueren, J. (2009). The Pragmatics of Interaction. John Benjamins Publishing.

Holt, E., & Clift, R. (2006). Reporting Talk : Reported Speech in Interaction. Cambridge University Press.

Poudat, C., Grabar, N., Paloque-Berges, C., Chanier, T. & Kun, J. (2017). « Wikiconflits : un corpus de discussions éditoriales conflictuelles du Wikipédia francophone » in Wigham, C.R & Ledegen, G., *Corpus de communication médiée par les réseaux : construction, structuration, analyse*. Collection Humanités numériques. Paris : L'Harmattan.

Poudat, C. & Ho-Dac, L-M. (2019). Désaccords et conflits dans le Wikipédia francophone. In Col, G. et Hanote, S. (Eds.), *Accord et désaccord*, Travaux linguistiques du Cerlico, 29: 155-176.

Rabatel, A. (2004). L'effacement énonciatif dans les discours rapportés et ses effets pragmatiques. *Langages*, 38(1 56), 3-17.

Rosier, L. (2002). La presse et les modalités du discours rapporté : L'effet d'hyperréalisme du discours direct surmarqué. *L'information grammaticale*, 94(1), 27-32.

Rosier, L. (2008). Le discours rapporté en français. Ophrys.

Corpus d'apprenants. Applications au-delà des théories de l'acquisition des langues.

Minerva Rojas [1]
[1] Laboratoire Bases Corpus Langage, UMR 7320. Université Côte d'Azur | CNRS
minerva.ROJAS@univ-cotedazur.fr

Introduction

La place de corpus de référence dans l'enseignement des langues s'avère indiscutable car ils constituent des données authentiques (Boulton, 2008 ; McEnery, Xiao, 2011) représentatives de la langue cible, permettant à l'apprenant de se rapprocher à différents usages, régularités et exceptions (Osborne, 2002). Ainsi, les corpus sont utilisés de manière directe guidant l'enseignement basé sur les données (*data-driven learning*) ou indirecte comme source de création de matériels pédagogiques et de choix des contenus (Römer, 2009). En effet, l'enseignement des langues secondes et l'étude de corpus de référence ont un lien étroit à partir des années 1950 (Kennedy, 1992). Par exemple, les contenus lexicaux et grammaticaux du Français fondamental (Gougenheim et al., 1964) ont été délimités à la suite de l'analyse de fréquence d'un corpus de référence de plus de 800 000 mots.

Or, toujours du point de vue pédagogique, les corpus d'apprenants ont été exploités plus tardivement et pendant longtemps ils ont constitué une sorte de « chainon manquant » entre la linguistique de corpus et la didactique des langues (Gilquin *et al*., 2009). Certes, les corpus d'apprenants occupent une place pertinente et bien justifiée dans la construction des théories de l'acquisition des langues (Granger, 2009) car l'analyse des productions langagières aide à tracer et modéliser le développement de la L2 (Freeman, Cameron, 2008 ; Verspoor, de Bot, Lowie, 2011 ; De Cock, Tyne, 2014). Pourtant ce serait seulement à partir de la fin des années 2000 que l'on trouve des travaux se penchant sur l'exploitation de corpus d'apprenants en enseignement des langues (Gilquin *et al*. 2009 ; Mas, Gil, 2018 ; Meunier, 2010).

Comme Meunier (2012) le signale, une des raisons expliquant la carence d'applications pédagogiques des corpus d'apprenants pourraient être dues à la méconnaissance sur l'existence ce type de corpus de la part des enseignants. Donc, la question se pose au moment de placer l'utilisation de corpus d'apprenants dans la formation des futurs enseignants (O'Keffer, Farr, 2003). En accord avec Xi (2017) leur utilisation dans l'évaluation de la production en L2 constituerait une opportunité pour fournir un retour détaillé aux apprenants sur leur performance. Dans ce cadre, la formation des futurs enseignants à l'exploitation des corpus lors de l'évaluation en L2 se révèlerait pertinente.

De plus, ce type de formation pourrait non seulement bénéficier aux futurs enseignants, mais l'enseignant-chercheur pourrait engager une révision théorique des variables analysées (Jarvis, 2017 ; Xi, 2017) grâce à la comparaison entre l'évaluation qualitative et quantitative menées avec les futurs enseignants. En outre, l'analyse quantitative d'un corpus (oral ou écrit) d'apprenants peut compléter l'évaluation qualitative basée sur des échelles telles que celles proposées dans le CECRL (Cushing, 2017). Par ailleurs ce type d'échelles peuvent comporter des biais lors de leur application par plusieurs évaluateurs, ce qui affecte leur fiabilité en tant qu'instrument de mesure (Weir, 2005). De plus, l'évaluation basée sur des échelles pourrait

déclencher un débat autour de l'opérationnalité de certains descripteurs du CECRL (Conseil de l'Europe, 2001, 2018), car leur rédaction n'a pas pris en compte des paramètres quantitatifs de la performance des apprenants mais le jugement des enseignants des langues (Conseil de l'Europe, 2018 : 45).

Pour ces raisons, nous avons mis en place une formation à l'exploitation quantitative et qualitative de corpus pour les futurs enseignants de français langue étrangère (FLE). Les objectifs étaient, en plus de la formation elle-même, i) de tester si l'évaluation qualitative utilisant les échelles du CECRL est fiable en mesurant le degré d'accord des juges ; ii) si l'évaluation qualitative peut saisir le développement interlangue des apprenants de FLE de la même manière que l'évaluation quantitative. Les résultats nous ont également permis d'entamer une discussion avec les futurs enseignants sur la faisabilité, les limites et les avantages de cette double approche de l'évaluation dans le contexte quotidien de la vie des enseignants.

Corpus

Pour ce travail nous comptons sur un corpus d'apprenants créé et exploité au cours de notre recherche doctorale (Rojas, 2020). Le corpus recueille les productions monologiques et dialogiques de 12 apprenants de FLE en immersion suivis pendant une période de 30 mois. Les productions orales ont été recueillies en quatre collectes de données (novembre 2015 : t1 ; mars, 2016 : t2 ; mars 2017 : t3 ; et novembre 2017 : t4).

Pour la recherche-formation présentée ici, le corpus a été échantillonné de manière aléatoire, ce qui a permis d'obtenir un échantillon de 12 productions orales monologiques de 6 locuteurs de FLE, dont 6 productions issues de la première collecte de données (t1) et 6 productions issues de la quatrième collecte de données (t4). Donc à chaque locuteur correspondent 2 productions, ce qui signifie un écart de deux ans entre chaque production. Ensuite les productions ont été divisées en deux lots et randomisées pour éviter un éventuel biais lors de leur évaluation qualitative : le premier lot est composé de 3 productions de t1 et de 3 productions de t4, et le deuxième lot est constitué de 3 productions de t1 et de 3 productions de t4.

Méthodologie

Cette démarche comporte deux étapes, l'une de formation et application de l'évaluation qualitative et l'autre de l'évaluation quantitative.

L'évaluation qualitative a été effectuée par un groupe d'étudiants de Master 2 FLE d'Université Côte d'Azur (n=7) à la suite d'une période de formation visant l'utilisation des échelles du CECRL pour l'évaluation de la production orale. Les échelles abordées dans la formation sont celles qui correspondent à la production orale générale, l'étendue de vocabulaire, la correction et l'aisance à l'oral (Conseil de l'Europe, 2001, 2018). Pour leur application, en plus du corpus oral, les étudiants ont reçu une grille d'évaluation contenant les quatre échelles afin d'attribuer à chaque locuteur de FLE un des six niveaux proposés dans le CECRL, à savoir : A1, A2, B1, B2, C1, C2 (Conseil de l'Europe, 2001).

Il faut signaler que l'évaluation qualitative a été réalisée en deux phases, avec un intervalle d'un mois entre chaque évaluation. Dans chaque phase les étudiants ont évalué un lot de productions. À la suite de chaque évaluation, les grilles remplies nous ont été envoyées sous format numérique afin de faciliter la collecte et l'analyse des données. Les résultats ont été soumis à des calculs de distribution puis à l'analyse du coefficient kappa de Fleiss (Fleiss,

1971). Ce coefficient détermine le degré d'accord entre les juges et si l'accord est statistiquement significatif ou non. Les Kappas peuvent être très faibles (0 à 0,20), faibles (0,21 à 0,40), modérés (0,41 à 0,60), forts (0,61 à 0,80) et presque parfaits (0,81 à 1,00).

Une fois l'évaluation qualitative réalisée, la formation à l'évaluation quantitative a eu lieu. Pour ce faire, il a été expliqué aux étudiants comment créer et exploiter un corpus annoté d'apprenants de FLE, comment lancer des recherches sur un concordancier et quels types de mesures quantitatives seraient pertinentes d'analyser. Ainsi, les unités d'analyse quantitatives que nous avons retenues sont l'indice D de diversité lexicale (ou VocD) (Jarvis, 2002), très utilisée dans les études en acquisition des langues (David, 2008 ; Hilton, 2014), le taux d'erreurs en relation avec la correction ou la précision (Laufer, Nation, 1995) et la longueur moyenne des segments (LMS), également une mesure classique de l'étude de la fluence (ou aisance à l'oral) en L2 (Towell *et al*., 1996 ; Segalowitz, 2010).

L'indice D se calcule avec le logiciel CLAN qui produit un coefficient allant de 0 à 125 ; plus le score est élevé, plus le lexique utilisé est riche et divers. Ce coefficient est issu d'une formule corrigée du TTR (*type token ratio*) (McCarthy, Jarvis, 2007). Le taux d'erreurs se calcule en collectant le nombre d'erreurs de chaque production, puis le divisant par le nombre de mots de la production analysée et multiplié ensuite par 100 o par 1000 (Hilton, 2014). Enfin, la LMS représente le nombre moyen de mots qu'un locuteur produit entre deux pauses de plus de 250 (Segalowitz, 2010). Les résultats des variables quantitatives ont également été soumis à des analyses statistiques descriptives et inférentielles (test de Wilcoxon) afin de décrire les tendances centrales et la dispersion des résultats dans l'ensemble du groupe et de savoir si les différences entre les résultats obtenus en t1 et t4 sont statistiquement significatives. En outre, ces trois variables quantitatives pourraient recouvrir trois des quatre échelles du CERCL évaluées : l'indice D recouvrirait l'étendue de vocabulaire ; le taux d'erreurs correspondrait à la correction ; et la LMS correspondrait à l'aisance.

En termes de résultats, pour ce qui est de l'évaluation qualitative, nous nous attendons à que les productions t4 soient positionnées à des niveaux plus élevés que les productions t1, et que les évaluateurs soient d'accord sur le jugement des productions sur chacune des quatre échelles du CECRL (Conseil de l'Europe, 2018). D'autre part, en termes de résultats quantitatifs, une augmentation de la diversité lexicale (indice D) et de la LMS est attendue, mais une diminution du taux d'erreurs.

Résultats

Pour rendre compte des résultats, nous distinguons aussi entre les résultats de l'évaluation qualitative et ceux de l'évaluation quantitative.

- En ce qui concerne l'évaluation qualitative, la distribution des niveaux attribués dans l'échelle de production orale générale (Figure 1) montre que les productions orales de t4 sont jugées comme ayant des niveaux plus élevés que celles de t1. En d'autres termes, de manière générale, l'évaluation qualitative réalisée par les mêmes évaluateurs rend compte de l'évolution de l'interlangue des locuteurs FLE.

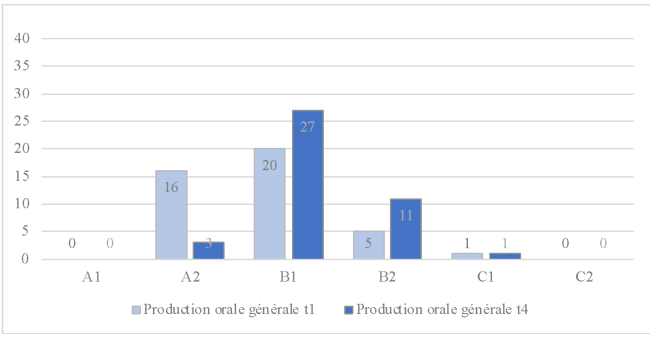


Figure 1. Distribution des jugements de la POG de t1 (2015) et de t4 (2017)

- Cependant, sur aucune des échelles évaluées, il n'y a eu un accord total entre les juges et, selon les résultats du coefficient kappa (table 1), leur degré d'accord n'a jamais été fort, mais plutôt faible ou modéré, sauf dans le cas de l'aisance à l'oral en t4.

	Kappa	Asymptotique			Intervalle de confiance asymptotique à 95 %	
		Erreur standard	z	Sig.	Limite inférieure	Limite supérieure
Production orale générale t1	.029	.064	.448	.654	-.097	.155
Production orale générale t4	.042	.067	.632	.527	-.088	.173
Étendue vocabulaire t1	.062	.180	.347	.729	-.291	.416
Étendue vocabulaire t4	.364	.175	2.073	.038	.020	.707
Aisance t1	-.012	.182	-.069	.945	-.369	.344
Aisance t4	.613	.179	3.428	.001	.262	.963
Correction t1	-.068	.194	-.350	.726	-.447	.312
Correction t4	.110	.162	.679	.497	-.207	.427

table 10. : **Résultats kappa de Fleiss. Degré d'accord entre évaluatrices**

- Les résultats quantitatifs (table 2) montrent que, dans l'ensemble du groupe, il y a une augmentation des valeurs de la moyenne de la LMS et de la diversité lexicale, même si les résultats du test Wilcoxon ne sont pas statistiquement significatifs. La moyenne du taux d'erreurs diminue dans l'ensemble du groupe, et les résultats ne sont pas non plus statistiquement significatifs.

							Stat. descriptives			**Wilcoxon**	
	Mg	Sr	Mt	Pt	Ya	Ua	N	M	SD	Z	Sig. Asint. (2-queues)
LMS t1	6.41	6.00	6.09	6.37	3.28	2.37	6	5.086	1.782		
LMS t4	7.10	4.89	5.36	6.4	5.42	5.63	6	5.800	.805	-.73	.463
Diversité lexicale t1	58.52	31.32	34.01	61.14	49.10	58.78	6	48.811	13.19		
Diversité lexicale t4	53.36	77.84	57.25	63.55	94.11	46.54	6	65.441	17.60	-1.15	.249
Taux d'erreurs t1	18.39	8.54	4.52	17.59	5.42	5.92	6	10.05	5.74		
Taux d'erreurs t4	7.33	9.62	4.41	15.53	15.20	6.30	6	9.731	4.2	-.08	.937

table 11. : **Résultats descriptifs et du test Wilcoxon des mesures quantitatives**

En résumé, l'évaluation qualitative peut saisir l'évolution du niveau en FLE de manière générale mais le degré d'accord entre les juges reste faible. Les résultats quantitatifs sont plus précis et, bien que non significatifs sur le plan statistique, ils peuvent compléter les jugements qualitatifs des évaluateurs. Dans notre communication, nous examinerons plus en détail les résultats, en nous concentrant sur les résultats individuels et les risques liés à leur généralisation. Nous évoquerons également les impressions des futures enseignantes sur la faisabilité de cette démarche dans leur pratique professionnelle.

Références bibliographiques

Boulton, A. (2008). Esprit de corpus : Promouvoir l'exploitation de corpus en apprentissage des langues. *Texte et corpus 3*, 37-46.

Conseil de l'Europe (2001). *Un cadre européen commun de référence pour les langues : Apprendre, enseigner, évaluer.* Strasbourg : Conseil de l'Europe.

Conseil de l'Europe (2018). *Un cadre européen commun de référence pour les langues : Apprendre, enseigner, évaluer. Un volume complémentaire avec des nouveaux descripteurs.* Strasbourg : Conseil de l'Europe.

Cushing, S.T. (2017). Corpus linguistics in language testing research. *Language Testing, 34*(4), 441-449.

De Cock, S., Tyne, H. (2014). Corpus d'apprenants et acquisition des langues. *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle, 11*(1). [En ligne : 10.4000/rdlc.1716]

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378-382.

Freeman, D. L., Cameron, L. (2008). Research methodology on language development from a complex systems perspective. *The Modern Language Journal, 92*(2), 200-213.

Gilquin, G., Granger, S., Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes, 6*(4), 319-335.

Gougenheim, G., Rivenc, P., Michéa, R., Sauvageot, A. (1964). *L'élaboration du français fondamental (1er degré) : étude sur l'établissement d'un vocabulaire et d'une grammaire de base (Vol. 2)*. Didier.

Hilton, H. (2014) Oral fluency and spoken proficiency: Considerations for research and testing. In A. Edmonds, P. Leclerq, H. Hilton. *Measuring L2 proficiency: Perspectives from SLA*, (pp. 27- 53). Berilin, New York: de Gruyter.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57-84.

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing 34*(4), 537-553.

Mas Álvarez, I., Gil Martínez, A. (2018). Los corpus de aprendices: un terreno en expansión para la enseñanza de español. Dans M. Ellison, M. Anido, P. Nicolás, S. Valente-Rodrigues.

As linguas estrangeiras no ensino superior:propostas didáticas e casos em estudo. (pp. 35-55). Porto : Universidade do Porto. Faculdade de Letras.

McCarthy, P. M., Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24* (4), 459-488.

McEnery, T., Xiao, R. (2011). What corpora can offer in language teaching and learning. In E. Hinkel (dir.) *Handbook of research in second language teaching and learning* (pp. 382-398). New York: Routledge.

Meunier, (2012) Learner corpora in the classroom: a useful and sustainable didactic resource. In L. Pedrazzini, A. Nava (dir.) *Learning and Teaching English: Insights from Research.* (p. 211-228). Milano: Poliletrica.

O'Keeffe, A., Farr, F. (2003). Using Language Corpora in Initial Teacher Education: Pedagogic Issues and Practical Applications. *TESOL Quarterly, 37*(3), 389–418.

Osborne, J. (2002). Integrating corpora into a language-learning syllabus. Dans B. Lewandowska-Tomaszczyk (dir.) *PALC 2001: Practical applications in language corpora* (pp. 479–492). Frankfurt: Peter Lang.

Rojas Madrazo, M. (2020). *Stratégies de communication, fluidité et lexique en production orale. Étude longitudinale d'apprenants de FLE en immersion*. Thèse de doctorat. Université Savoie Mont Blanc. Chambéry.

Römer, U. (2009). Corpora and language teaching. In A. Lüdeling, M. Kytö (Eds.) *Corpus Linguistics. An International Handbook* (pp.112-131). Berilin, New York: de Gruyter.

Segalowitz, N. (2010) *Cognitive bases of second language fluency*. New York: Routledge.

Towell, R., Hawkins, R., Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied linguistics*, *17*(1), 84-119.

Verspoor, M., de Bot, K., Lowie, W. (2011). *A dynamic approach to second language development: Methods and techniques* (Vol. 29). Amsterdam, Philadelphia: John Benjamins Publishing.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language testing, 22*(3), 281-300.

Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Language Testing 34*(4), 565-577.

(Re)catégoriser les connecteurs par l'étude de leur entourage dans des corpus de différents genres[98]

Corinne Rossari[1], Cyrielle Montrichard[1] et Claudia Ricci[1]
[1] Chaire de linguistique française, Université de Neuchâtel
corinne.rossari@unine.ch, cyrielle.montrichard@unine.ch, claudia.ricci@unine.ch

Introduction

Les connecteurs sont en général catégorisés en fonction de leurs propriétés relationnelles, qu'elles soient de nature argumentative, logique ou inférentielle, comme on peut l'observer dans de nombreux manuels de référence (voir par exemple Dubois *et al.,* 2018, Abeillé & Godard, 2021). Les études, surtout dans la littérature francophone, ont cherché à départager les connecteurs véhiculant un même type de relation; depuis l'article fondateur sur *car*, *parce que*, *puisque* du groupe Lambda-L (1975), plusieurs travaux ont suivi cette perspective visant à faire ressortir des propriétés différenciant des connecteurs ou des paires de connecteurs très proches sémantiquement (Anscombre & Ducrot, 1977 ; Rossari 1994 ; Guimier, 2000 ; Anscombre, 2002 ; Mellet & Monte, 2005), en mettant au cœur de leurs études des contextes qui manifestent leur compatibilité ou incompatibilité. Des exemples comme *Je n'aime pas les maths mais/?pourtant j'aime le français* ou *je n'aime pas le fromage alors que/?tandis que j'aime la charcuterie* permettent de pointer des différences dans l'usage de connecteurs sémantiquement apparentés.

Notre propos vise à participer à l'étude des caractéristiques qui rassemblent et différencient les connecteurs en interrogeant des corpus textuels s'inscrivant dans différents genres au moyen d'une approche issue de la statistique textuelle (Lebart & Salem, 1994). Il s'agira, dans un premier temps, d'observer le comportement d'une série de connecteurs relevant de différentes catégories par une étude fine de leur cotexte en interrogeant leurs combinaisons statistiquement significatives avec des catégories morphosyntaxiques peu sensibles aux variations de genres (les adverbes, les conjonctions, les verbes cognitifs et de modalité). Cette première étape révélera que la proximité qu'on observe entre certains connecteurs ne suit pas forcément leur proximité sémantique. Nous proposerons, dans un deuxième temps, de nous focaliser sur la catégorie spécifique des connecteurs de concession afin d'interroger plus finement ce qui permet de distinguer, dans leur usage, des connecteurs proches sémantiquement. Nos investigations prendront en compte deux langues romanes, le français et l'italien, pour évaluer la prégnance du code sur les principes cognitifs régissant les relations véhiculées par les connecteurs.

Corpus et méthodologie

Notre méthodologie doit donc permettre d'interroger statistiquement des textes en français et en italien dans une perspective contrastive en genre de discours.

[98] Ce travail s'inscrit dans le cadre du financement d'un projet FNS intitulé « L'ancrage argumentatif des formes modales. Etude sur corpus avec un éclairage comparatif entre français et italien », projet n. 100001F_192247.

Ainsi, les corpus sur lesquels l'étude se fonde sont constitués de textes écrits relevant de différents genres discursifs qui varient au niveau de leur dimension argumentative et subjective. Nous interrogeons donc trois types de discours : (i) un ensemble de discours politiques, corpus qui comporte un fort degré de subjectivité et d'argumentation, (ii) un ensemble de discours de presse, qui comporte un degré intermédiaire de subjectivité et d'argumentation et (iii) un discours encyclopédique, qui revendique une neutralité au niveau de la subjectivité et de l'argumentation.

Corpus	Date de publication	Langue	Genres discursif	Nombre d'occurrences
Discours de présidents français	(1958-2020)	Français	Discours politiques	2'843'577
Discours de présidents italiens	(1955-2021)	Italien	Discours politique	2'639'007
Est Républicain	2010	Français	Presse écrite régionale	18'669'845
Adige	1999-2006	Italien	Presse écrite régionale	20'343'148
Wikipédia	2019	Français	Encyclopédie	18'715'455
Wikipédia	2019	Italien	Encyclopédie	18'394'987

table 1. Description des corpus textuels sélectionnés pour l'étude

Pour exploiter et interroger ces données textuelles nous mobilisons trois outils statistiques issus de la textométrie, approche qui permet d'allier étude quantitative et qualitative :

- (i) Le calcul de la cooccurrence spécifique (au moyen du logiciel TXM et du logiciel R), qui permet d'interroger les rapports privilégiés qu'entretient chaque connecteur avec les formes indiquant un degré de subjectivité particulier[99].
- (ii) L'Analyse Factorielle de Correspondances (AFC), fondée sur la co-fréquence entre un paradigme de connecteurs et des formes préalablement définies – adverbes entre autres (Mayaffre, 2014 ; Viprey, 2016). Ce calcul sera effectué au moyen d'Hyperbase, logiciel développé par Brunet (2012). La représentation graphique permettra de faire contraster le profil des connecteurs selon les formes qui leur sont statistiquement associées.
- (iii) L'analyse arborée (Mayaffre & Luong, 2003) permettra quant à elle de mesurer la distance entre chaque profil de connecteur afin d'observer la proximité ou la distance entre connecteurs selon leur co-fréquence avec la liste prédéfinie.

Les trois calculs mobilisés permettent d'adopter une granularité plus ou moins fine et différents angles de lecture selon les formes interrogées. En effet, la recherche des cooccurrents intègre toutes les formes spécifiques appartenant à une catégorie

[99]Les indices sont mesurés à l'aide du Log-Likelihood (LL) qui s'interprète comme suit : tous les indices supérieurs à 10,83 sont le signe d'une attraction significative entre deux items dans un empan donné. Après plusieurs tests, nous avons opté pour un empan interrogeant 10 items avant le connecteur et 5 après afin de saisir les adverbes en position initiale.

morphosyntaxique (les conjonctions ou les adverbes par exemple) alors que l'AFC et l'analyse arborée se fondent sur une liste de formes préalablement définie, construite selon des critères formels, sémantiques et de fréquence. Ainsi, la méthodologie que nous utilisons est fondée d'une part sur le croisement de ces trois méthodes statistiques et d'autre part sur le degré de granularité des données qu'elles permettent de présenter. D'abord, pour la méthode (i), nous utilisons des calculs fondés sur des indices de spécificité en relation avec des macro-catégories de formes. Nous avons débuté cette recherche par les adverbes, en distinguant, d'après la classification de Molinier et Levrier (2000) les adverbes modaux et disjonctifs de style – qui, selon nous, ont la propriété de mettre explicitement en scène la voix du locuteur – des adverbes de manière, qui, tout en communiquant pour la plupart d'entre eux une évaluation imputable au locuteur, ne comportent pas la voix de ce dernier comme une composante de leur sémantique (cf. par exemple, la différence entre *franchement* et *lentement*). En effet, un adverbe modal ou qualifiant l'énonciation est une prise de position qui est par défaut imputée au locuteur sur la proposition, d'où le fait qu'ils sont généralement qualifiés comme des adverbes d'attitude ou d'énonciation. En revanche, un adverbe de manière est partie intégrante du prédicat verbal et ne communique pas directement une appréciation du locuteur sur l'état de choses[100]. (ii) Ensuite, pour l'AFC de cooccurrences, nous avons mobilisé des méthodes fondées sur des co-fréquences avec des formes précises, ce qui permet d'interroger les liens privilégiés que les connecteurs entretiennent avec une sélection d'adverbes. (iii) Enfin, l'analyse arborée est utilisée pour mettre la focale sur la distance entre les connecteurs dans leur relation avec les adverbes.

Résultats

Nous avons établi des recherches fondées sur les trois méthodes statistiques concernant une série de connecteurs signalant quatre types de relations discursives : temporelles, causales, consécutives et oppositives. Nous présentons ici les résultats de nos recherches préliminaires concernant deux des trois corpus pris en compte : le corpus de presse et le corpus encyclopédique. Les indices de spécificité relevés pour les deux macro-catégories d'adverbes mettent en relief une différence entre : i. des connecteurs attirés par les deux catégories ; ii. des connecteurs attirés uniquement par la catégorie d'adverbes de manière et iii. des connecteurs qui ne sont attirés par aucune des deux catégories. Ces recherches montrent des constantes à la fois inter-langues et inter-genres : les connecteurs *mais* et *ma* sont les plus fortement attirés par les deux catégories dans tous les genres, avec des différences inter-langues – *ma* est davantage attiré par les adverbes énonciatifs alors que *mais* est davantage attiré par les adverbes de manière – et des différences inter-genres – *alors que* est attiré par les deux catégories d'adverbes dans le corpus de presse et est attiré uniquement, et dans une moindre mesure, par les adverbes de manière dans le corpus encyclopédique (les figures 1 à 4 ci-après montrent les résultats mentionnés ici). Ces premiers résultats seront mis en perspective avec les deux autres méthodes, afin d'observer dans quelle mesure le comportement des connecteurs peut être considéré comme stable. Par exemple, si une différence nette entre connecteurs révélée par les indices de cooccurrents spécifiques s'observe

[100] Pour tous les adverbes pouvant être associés à l'une ou l'autre des catégories (comme *simplement* par exemple), nous procéderons en deux étapes : (i) un tri automatique fondé sur leur position syntaxique vérifié manuellement sur des échantillons – *simplement* suivi d'une virgule en début de phrase prend la fonction de disjonctif de style alors que quand il est directement positionné après le verbe ou l'auxiliaire il prend la fonction d'adverbe de manière –, (ii) un tri manuel pour les formes qui ne peuvent pas être désambiguïsées par leur position syntaxique.

également dans les deux autres méthodes, nous considérons que le comportement de ceux-ci est stable - dans la mesure où les données mobilisées dans chaque méthode sont sensiblement différentes. Les points d'instabilité pourront ainsi être repérés et faire l'objets de nouvelles analyses.

Notre étude propose ensuite de se concentrer sur une catégorie spécifique des connecteurs : les concessifs. Nous présentons ici une AFC de cooccurrences menée sur le corpus de presse *L'Est Républicain* 2010 (voir figure 5). Cette dernière met en évidence des différences de comportement très nettes par rapport aux deux catégories d'adverbes : *mais* et *alors que* sont entourés par des adverbes énonciatifs (*simplement, malheureusement, vraiment, évidemment*…) et s'opposent, sur l'axe des abscisses, à *cependant* et *pourtant*, qui montrent dans leur entourage des adverbes non énonciatifs (*directement, parfaitement, totalement, largement, particulièrement*…).

Notre contribution proposera de présenter et mettre en contraste les trois méthodes statistiques mobilisées et ce qu'elles permettent de mettre en lumière dans le fonctionnement de connecteurs sémantiquement proches. Notre approche quantitative sera appuyée par des remises en contexte et une analyse qualitative fine fondée sur des extraits issus des corpus interrogés afin d'illustrer des usages prototypiques des connecteurs selon le genre dans lequel ils sont employés.

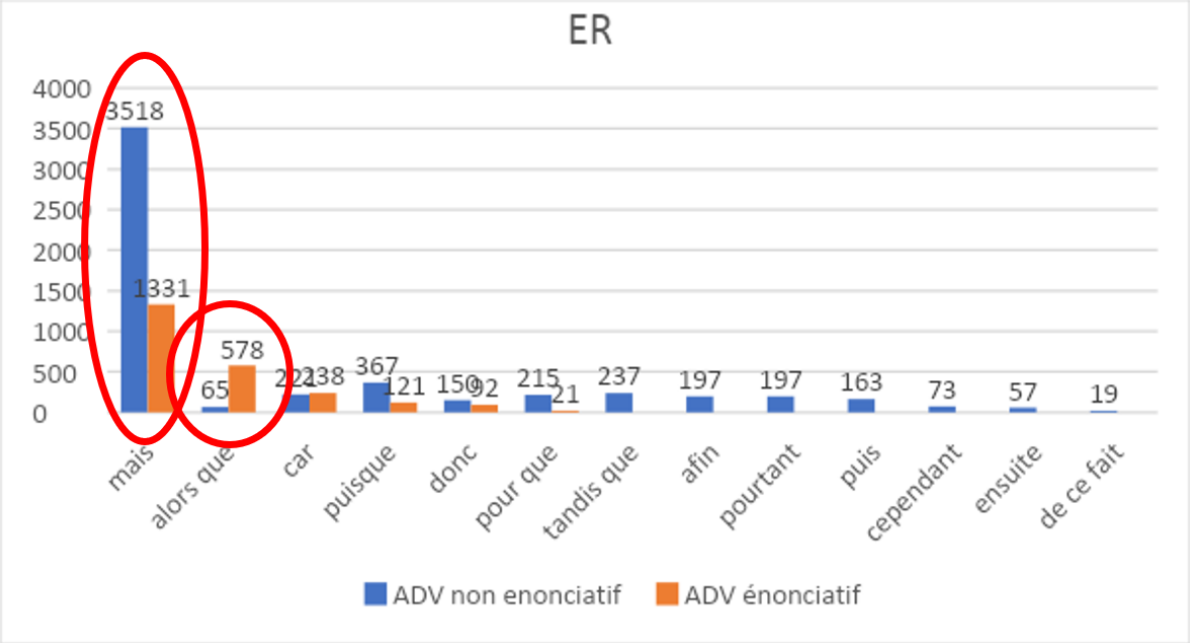


figure 1. Cooccurrence spécifique entre connecteurs et adverbes énonciatifs vs non-énonciatifs – corpus de presse L'Est Républicain (indice de LL)

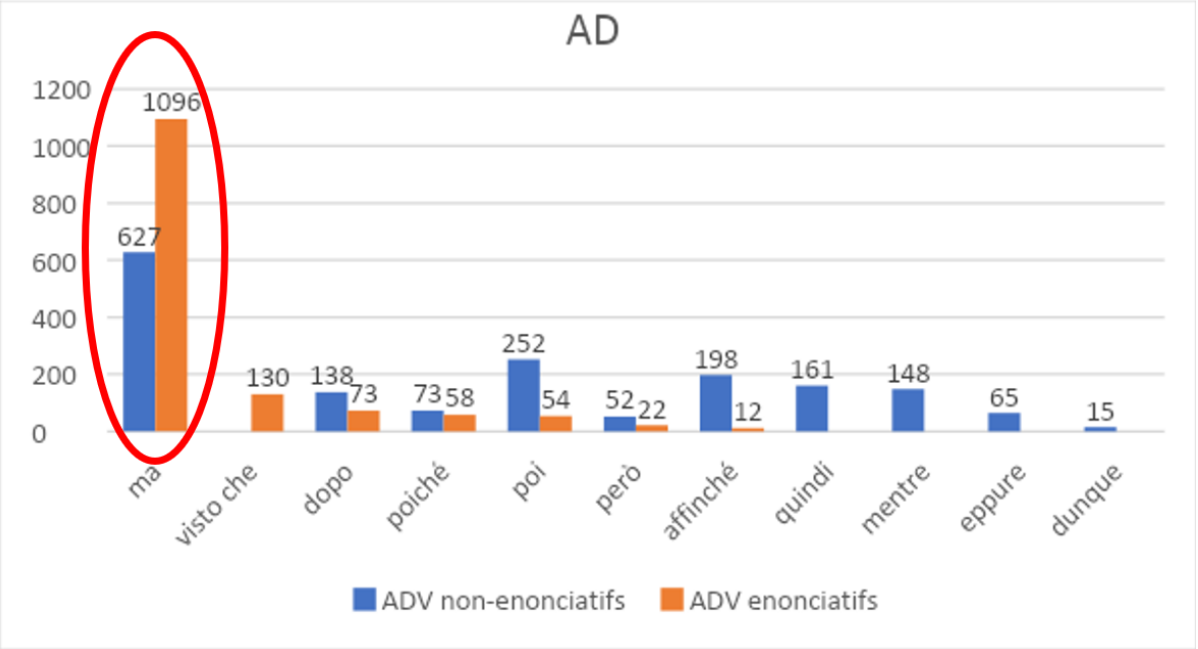


figure 2. Cooccurrence spécifique entre connecteurs et adverbes énonciatifs vs non-énonciatifs – corpus de presse L'Adige (indice de LL)

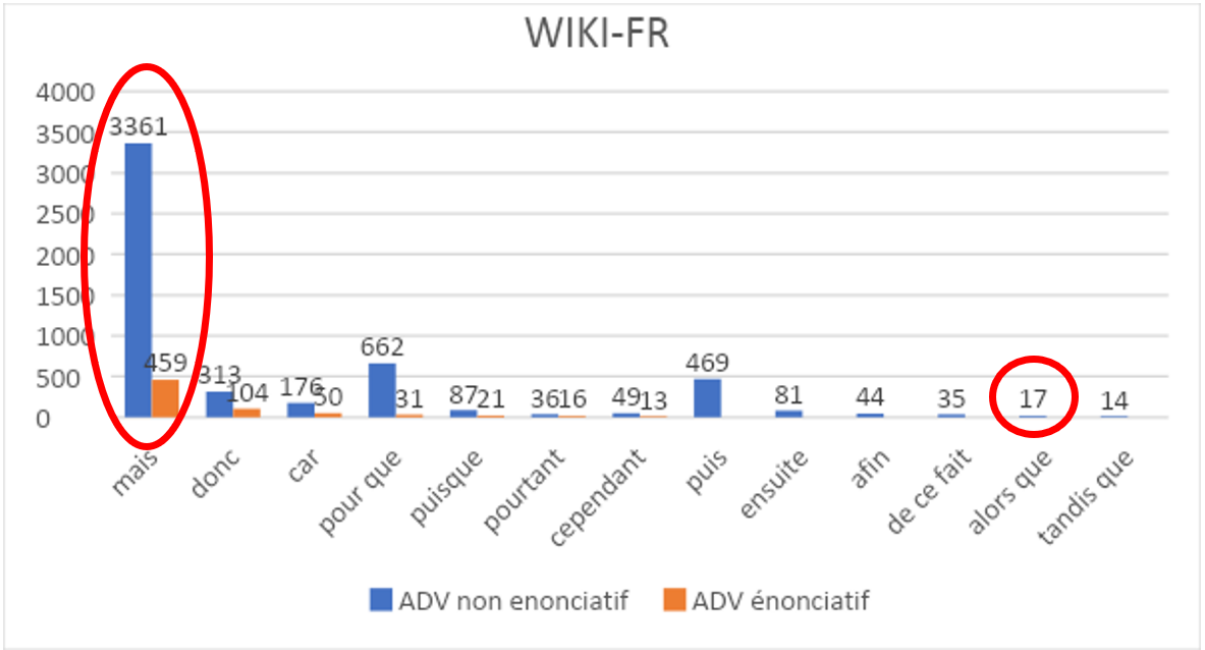


figure 3. Cooccurrence spécifique entre connecteurs et adverbes énonciatifs vs non-énonciatifs – corpus encyclopédique Wikipédia français (indice de LL)

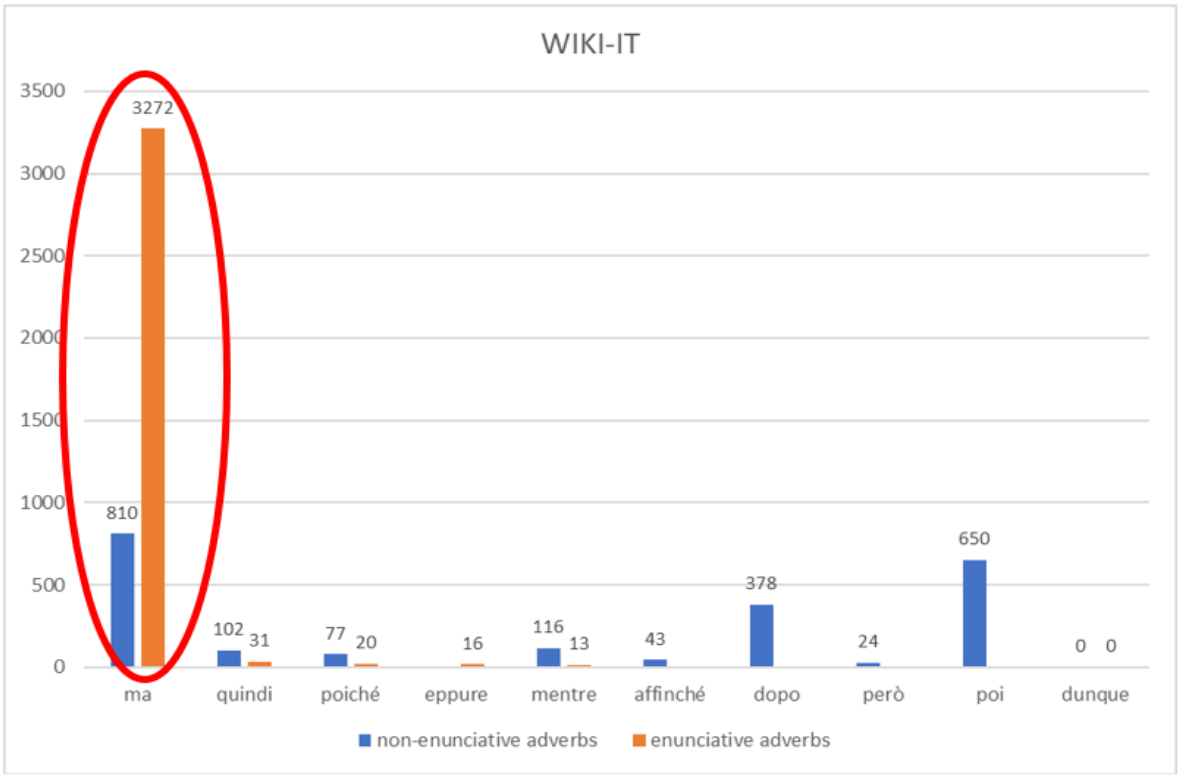


figure 4. Cooccurrence spécifique entre connecteurs et adverbes énonciatifs vs non-énonciatifs – corpus encyclopédique Wikipedia italien (indice de LL)

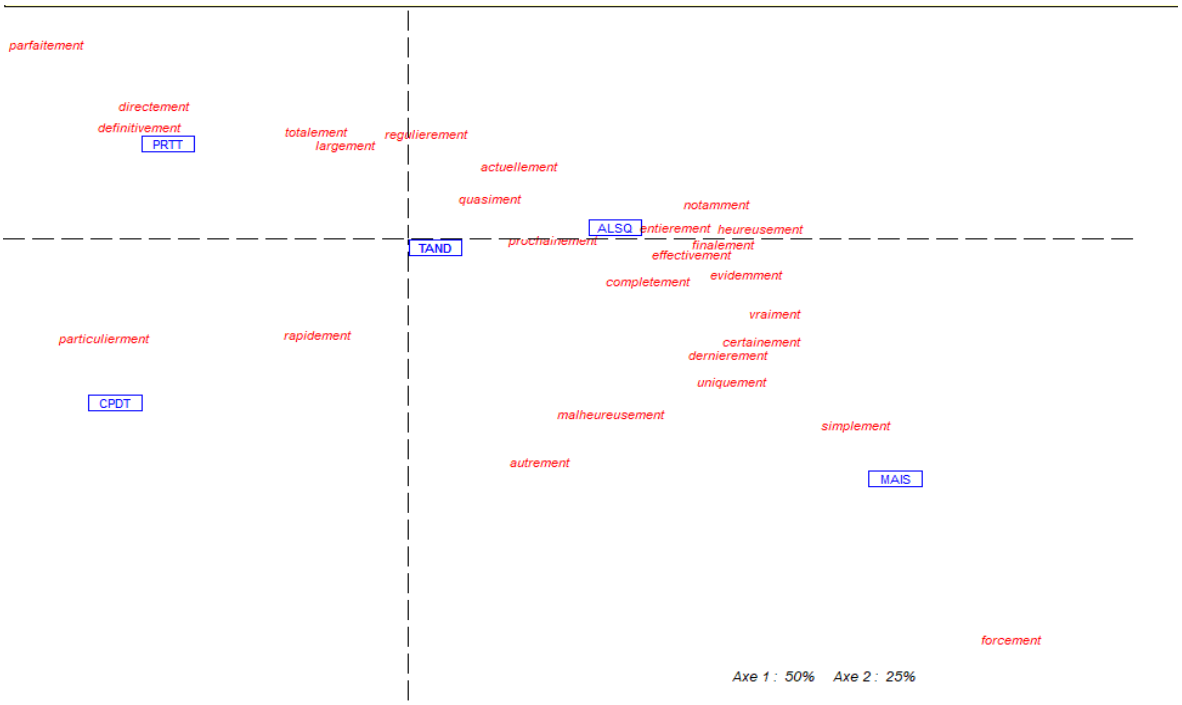


figure 5. AFC de cooccurrences sur les connecteurs (en bleu[101]) et adverbes (en rouge) – corpus de presse L'Est Républicain

[101] Légende pour les connecteurs abrégés : ALSQ = alors que ; CPDT = cependant ; PRTT = pourtant ; TAND = tandis que.

Références bibliographiques

Abeillé, A., Godard, D. (dirs) (2021). *La Grande Grammaire du français*. Paris, Arles : Actes Sud / Imprimerie Nationale.

Anscombre, J.-C. (2002). Mais/pourtant dans la contre-argumentation directe : raisonnement, généricité, et lexique. *Linx*, n.46, 115-131. URL : http://journals.openedition.org/linx/104

Anscombre, J.-C., Ducrot, O. (1977). Deux mais en français ?. *Lingua*, n.43, 23-40. URL : https://doi.org/10.1016/0024-3841(77)90046-8.

Brunet E. (2012). Nouveau traitement des cooccurrences dans Hyperbase. *Corpus* [en ligne], 11. URL: http:// journals.openedition.org/corpus/2275

Dubois, J., Giacomo, M., Guespin, L., Marcellesi, C., Marcellesi, J.-C. & Mével, J.-P. (2018). *Le dictionnaire de linguistique et des sciences du langage.* Paris : Larousse dictionnaires.

Groupe Lambda-l (1975). Car, parce que, puisque. *Revue Romane,* n.10, 248-280.

Guimier, C. (2000). Non-congruence et congruence : alors que vs tandis que. *Syntaxe et Sémantique*, n.1-1, 80-112. URL : https://doi.org/10.3917/ss.001.0080.

Heiden, S., Mague, J-Ph, Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie-conception et développement. Bolasco, S., Chiari, I. & Giuliano, L. (dirs). *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*. v. 2., n. 3. Milano : Edizioni Universitarie di Lettere Economia Diritto, 1021-1032. URL : https://halshs.archives-ouvertes.fr/ halshs-00549779.

Lebart, L., Salem, A. (1994). *Statistique textuelle.* Paris : Dunod.

Mayaffre, D. (2014). Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014). Née, E., Daube, J.-M., Valette, M., Fleury, S. (éds). . *Actes des JADT 2014 (12es Journées internationales d'analyse statistique des données textuelles)* Paris : Inalco-Sorbonne nouvelle, 15-32.

Mayaffre, D., Luong, X. (2003). Les discours de Jacques Chirac (1995-2002). *Histoire & mesure*, XVIII 3-4: 289-311.

Mellet, S., Monte, M. (2005). Néanmoins et toutefois : polyphonie ou dialogisme ?. Bres, J. et al. (éds). *Dialogisme et polyphonie*. Bruxelles : Duculot, 249-263.

Molinier, Ch., Levrier, F. (2000). *Grammaire des adverbes. Description des formes en -ment*. Genève/Paris : Droz.

Rossari, C. (1994). De *donc* à *dunque* et *quindi* : les connexions par raisonnement inférentiel. *Cahiers de linguistique française*, n.15, 7-49.

Viprey, J.-M. (2016). Comparer des AFC de cooccurrence généralisée. *Journées internationales d'analyse de données textuelles*. Jun 2016, Nice, France. URL: http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/83505/83505.pdf

Les apports des corpus numériques pour la formation des étudiants de Master FLE

Ruggia Simona
Laboratoire Bases, Corpus, Langage UMR 7320 UCA/CNRS/France Université Côte d'Azur
Simona.Ruggia@univ-cotedazur.fr

Introduction

Les recherches en didactique des langues ont mis au jour plusieurs usages et par conséquent rôles des corpus (Ruggia, Gaillat, dir. 2022). En 1993, Fligelstone en a distingué trois : le « teaching about » où le corpus est un « objet d'enseignement » (Cavalla, Loiseau, 2013 : 2), le « exploiting to teach » où le corpus correspond à un « support d'enseignement » (*Ib.*) et le « teaching to exploit » qui consiste à « exploiter les corpus et l'interprétation des résultats pour enseigner une langue » (*Ib.*). Un quatrième usage a ensuite été défini par Renouf (1997) : le « teaching to establish resources » pour former à la création de corpus. Ainsi, la notion de corpus s'est élargie dans la lignée de la linguistique de corpus à travers des expérimentations d'utilisation de corpus en classe de langue. Dans cette optique, le corpus est devenu un objet d'apprentissage pour l'apprenant, selon l'approche définie « *data-driven learning* » par Johns (1988 ; Johns & King, 1991) qui mobilise des stratégies de découverte en développant la méta-compétence « apprendre à apprendre » (Holec, 1990). Selon cette approche, transposée en français « apprentissage sur corpus » (ASC) et développée notamment par Boulton et Tyne (2014), « l'apprenant est amené à mettre à profit ses différentes observations de la langue à partir de données qui se présentent sous forme de corpus […] » (*Ib.* : 6). Mais comment enseigner et apprendre à l'aide d'un corpus ? Il va sans dire que l'utilisation de corpus nécessite des compétences « tant sur les plans théoriques et méthodologiques, que sur le plan technique » (Kübler, 2014 : 29), ce qui implique la mise en place de parcours adaptés pour la formation de futurs enseignants en fonction d'objectifs pédagogiques ciblés (Cavalla, 2019).

Méthodologie et corpus

Dans le cadre de nos travaux en didactique du français langue étrangère (FLE), nous nous sommes intéressée à l'étude outillée de corpus en tant qu'objet de recherche mais aussi en tant qu'objet et support d'enseignement/apprentissage. Ainsi, notre étude s'inscrit dans un projet de recherche que nous menons depuis quelques années sur la classification des niveaux selon les échelles du *Cadre Européen Commun de Référence pour les Langues* (CECRL) (Conseil de l'Europe, 2001, 2018) et sur les apports des corpus numériques pour l'évaluation, l'analyse et la description des spécificités des textes, autrement dit des « séquence[s] discursive[s] orale[s] et/ou écrite[s] » (Conseil de l'Europe, 2001 : 15) en fonction des 6 niveaux. Ces travaux, qui s'appuient sur une approche interdisciplinaire engageant un dialogue entre didactique du FLE, Intelligence Artificielle (IA) et Analyse des Données Textuelles (ADT), ont permis de réaliser la plateforme : DeepFLE (http://deeptext.unice.fr/FLE/) (Ruggia, Vanni, 2021) qui propose la classification de textes oraux via *deep learning* ainsi qu'une description des saillances apprises par l'IA qui marquent un changement de niveau. La création de cette plateforme a été possible grâce à la constitution d'un corpus numérique et échantillonné : le Corpus Manuels FLE qui est composé de textes oraux, (transcriptions d'interactions et

monologues, présents dans divers ensembles pédagogiques de FLE) et organisé en 6 classes correspondant aux 6 niveaux du CERCL (Ruggia, 2021). Il s'agit d'un corpus lemmatisé avec Spacy, homogène en taille et composé d'une masse critique de données pour chaque classe, soit 100 000 occurrences.

Résultats

Les résultats de ces travaux nous ont amenée à mettre en place un parcours réflexif de construction d'une compétence qui vise la maîtrise des spécificités lexicales et morpho-syntaxiques des séquences discursives orales en fonction des échelles des niveaux du CECRL. Ce parcours est destiné aux étudiants du Master FLE, dans le cadre de l'unité d'enseignement de didactique de l'oral, que nous assurons à Université Côte d'Azur. L'étude des niveaux est d'abord effectuée à travers l'analyse des descripteurs et des grilles d'évaluation du CECRL, des inventaires des *Référentiels pour le français* (Beacco *et al.* 2004a, 2007, 2008, 2011 ; Riba, 2016), et des vidéos de productions orales d'apprenants allophones, illustrant les niveaux du CECRL. Ces outils, certes incontournables, ont des limites qui ont été mises au jour par certains didacticiens. En ce qui concerne les descripteurs, qui présentent en termes positifs les compétences pour chaque type d'activité langagière, Tagliante a notamment précisé que « tels qu'ils sont formulés […] [ils] ne sont pas tous directement évaluables » (Tagliante, 2005 : 61), ce que confirment également les auteurs du CECRL lorsqu'ils affirment que « les échelles de descripteurs […] n'ont pas pour fonction première d'être destinées à l'évaluation » (Conseil de l'Europe, 2018 : 42). D'autre part, les *Référentiels pour le français* illustrent les contenus d'enseignement et par conséquent les spécificités de chaque niveau en termes de « réalisations linguistiques », mais force est de constater que ces ouvrages semblent rappeler les listes de vocabulaire encore très présentes dans les manuels de FLE et pourtant bannies (Tinkham, 1993) en faveur d'une étude en contexte. C'est justement pour favoriser une analyse en contexte des caractéristiques des niveaux que nous avons intégré l'exploration d'un corpus numérique, tel que le Corpus Manuel FLE, dans la formation des étudiants du Master FLE. Dans ce sens, l'IA et l'ADT enrichissent et complètent la description des niveaux, mais surtout en permettent l'évaluation : le *deep learning* grâce à son analyse qui est la fois évaluative et descriptive et l'ADT grâce à l'analyse statistique qui met au jour des observables linguistiques complexes susceptibles de caractériser un discours.

Dans cette approche, le corpus Manuels FLE est à la fois un objet et un support d'enseignement/apprentissage (Fligelstone, 1993 ; Cavalla, Loiseau, 2013 ; Boulton, Tyne, 2014). Un objet d'enseignement car la première étape : le « *teaching about* » consiste en l'enseignement de la linguistique de corpus que Kübler considère comme « une approche théorique et méthodologique qui, appliquée à l'apprentissage des langues, a pour objectif d'amener à se poser les bonnes questions pour obtenir les bonnes réponses » (2014 : 1). Dans la lignée de Kübler, nous estimons que cette approche possède la même valeur heuristique, appliquée à la formation des futurs enseignants de FLE. Le corpus est aussi un support d'enseignement « *exploiting to teach* », parce que son exploration, effectuée aussi bien selon l'approche *corpus-based* que *corpus-driven* (Tognini-Bonelli, 2001), permet à l'enseignant de « traiter par le corpus des questions lexicales, […] des questions de grammaire et des questions de contenus » (Kübler, 2014 : 28). Enfin, pour l'apprenant-étudiant, le corpus représente « un objet et un support d'apprentissage », les étudiants deviennent « les

utilisateurs des données » (Boulton, Tyne, 2014 : 7) et en quelque sorte des « chercheurs » (*Ib.*) qui observent, découvrent et s'informent sur la langue.

Dans le cadre de notre parcours, l'observation et la découverte des caractéristiques lexicales et morphosyntaxiques des niveaux du CECRL à travers l'exploration d'un corpus numérique, permet aux étudiants de maîtriser les niveaux.

L'étude du corpus est effectuée aussi bien à l'aide de la plateforme Hyperbase web (http://hyperbase.unice.fr/hyperbase/) que de la plateforme DeepFLE.

Hyperbase, qui combine des fonctions documentaires et statistiques, permet d'une part, l'enseignement de l'ADT et d'autre part l'exploration du corpus pour décrire, caractériser, classer et interpréter les textes.

Grâce à Hyperbase web le corpus est analysé d'un point de vue qualitatif et quantitatif. A titre d'exemple, la figure 1 illustre les résultats de la fonction « recherche » sous forme de concordancier de l'unité linguistique « pantalon » dans tout le corpus[102].

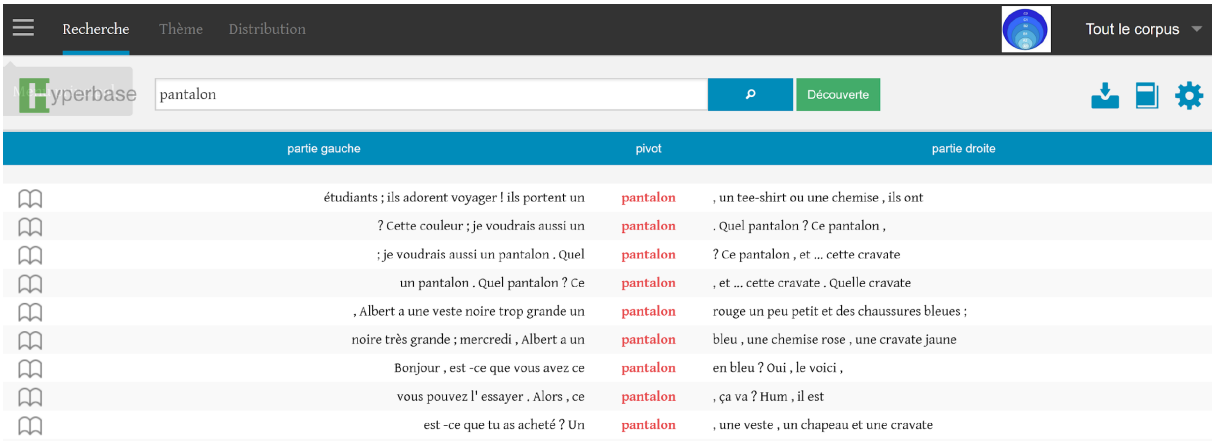


Figure 1. : Occurrence du pivot « pantalon » avec Hyperbase web

Cette fonction qualitative favorise une étude en contexte des unités linguistiques, d'abord dans leur contexte local, autrement dit avec le cotexte gauche et droit, et ensuite dans leur contexte global grâce au retour au discours en affichant le texte qui contient le pivot.

D'autre part, l'analyse quantitative des occurrences d'un pivot démontre de quel niveau du CECRL il est typique, puisqu'elle donne sa distribution dans les 6 classes du corpus :

[102] Cette figure ne reproduit que le début des résultats qu'on peut visualiser en faisant défiler la page.

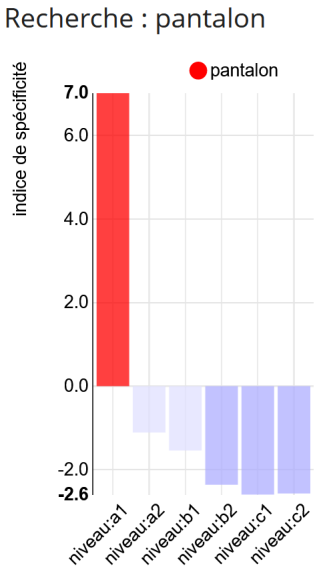


Figure 2. : Distribution fréquence des occurrences du pivot « pantalon » avec Hyperbase web

Dans ce sens, le graphique de la figure 2 prouve que le pivot « pantalon » est spécifique au niveau A1. Cette fonction d'Hyperbase est particulièrement intéressante pour étudier le lexique ainsi que les catégories morpho-syntaxiques qui caractérisent les niveaux du CECRL, tels qu'ils sont décrits dans le *Niveau B2 pour le français, textes et références* (Beacco, Bouquet et Porquier, 2004b : 33, 45-46, 59, 72-73, 84-85, 96). Cet ouvrage présente une liste de « objectifs communicatifs », « notions » et « catégories morpho-syntaxiques » pour chaque niveau dont les réalisations linguistiques sont illustrées dans les *Référentiels pour le français* de chaque niveau (Beacco *et al.* 2004a, 2007, 2008, 2011, Riba, 2016). En ce qui concerne l'exemple du pivot « pantalon » il est à noter que ce dernier appartient à la « notion : vêtement » ainsi qu'à « l'objectif communicatif : faire des achats ». Enfin, nous tenons à préciser que nous n'avons illustré *supra* qu'une requête à partir d'une simple « forme graphique » pour expliquer comment commencer à explorer le corpus avec les étudiants en Master FLE. L'exploration est ensuite complétée par d'autres requêtes qui peuvent combiner une « forme graphique » et/ou un « code grammatical » et/ou « un lemme », dont nous donnerons un exemple dans la figure 4.

L'étude des spécificités des niveaux des séquences discursives en fonction des niveaux du CECRL est aussi effectuée grâce à la plateforme DeepFLE. Cette dernière, qui est capable de prédire et détecter le(s) niveau(x) d'un texte en quelques millièmes de secondes, permet de découvrir et apprendre les passages-clés typiques de chaque niveau. Pour ce faire, il suffit de copier/coller un texte et d'en demander la détection.

Figure 3. Exemple de prédiction et description d'un texte avec DeepFLE

Comme l'illustre la figure 3, la prédiction du niveau est affichée aussi bien sous forme de diagramme type radar que de score de reconnaissance. En ce qui concerne l'analyse descriptive des passages-clés, les résultats sont visibles, grâce aux couleurs attribuées aux marqueurs linguistiques identifiés. De cette manière, le bleu indique qu'il s'agit d'une occurrence que le système reconnaît en tant que mot, l'orange indique la catégorie grammaticale et des codes précisent le type de la catégorie (DET:ART : articles …), et le vert signale qu'il s'agit d'un lemme. DeepFLE donne le niveau global d'un texte mais reconnaît aussi les passages qui peuvent correspondre à un niveau inférieur ou supérieur, en illustrant ainsi l'organisation enchâssée des niveaux qui « sont inclus les uns dans les autres » (Beacco *et al.* 2008 : 15). A ce propos, le texte de la figure 3 contient 57,14% de passages-clés du niveau A1 et 42,86% du niveau A2. Pour visualiser l'analyse de tous les passages-clés, l'utilisateur doit cliquer sur « précédent / suivant ».

L'étude des passages-clés spécifiques d'un niveau détectés par DeepFLE et plus particulièrement la distribution statistique des marqueurs le plus fortement attribués à une classe est ensuite analysée avec Hyperbase web. A titre d'exemple, l'analyse statistique de l'enchaînement syntaxique « je cherche un appartement situé » (figure 4) corrobore les résultats de la plateforme DeepFLE (figure 3).

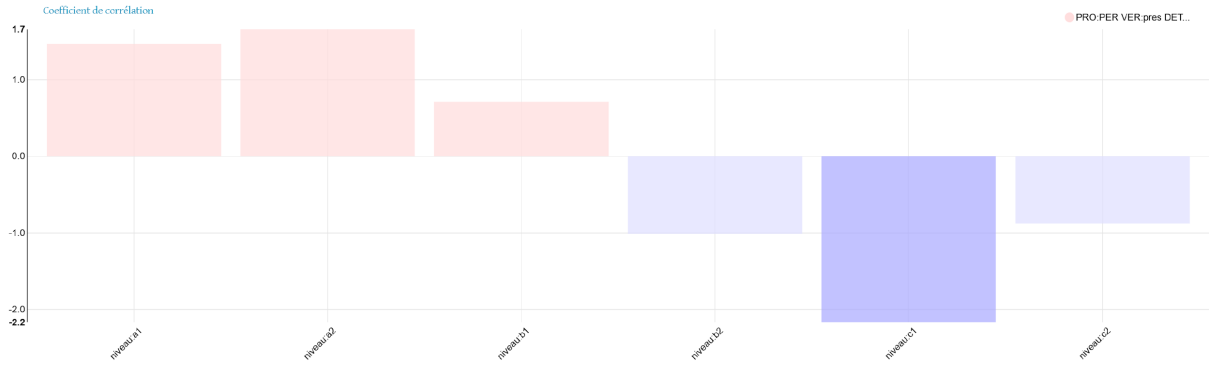


Figure. 4 : Distribution statistique d'un enchainement syntaxique avec Hyperbase web

Conclusion et perspectives

Notre contribution a mis au jour les apports des corpus numériques pour la formation des étudiants de Master FLE et plus précisément pour l'enseignement/apprentissage de compétences ciblées, à savoir la maîtrise et l'évaluation des spécificités de séquences discursives orales en fonction de leur niveau du CECRL. Ainsi, les quelques exemples d'analyses qualitatives et quantitatives que nous avons illustrés ont montré comment un corpus numérique tel que le Corpus Manuels FLE peut être appréhendé en tant qu'objet et support d'enseignement/apprentissage.

Par ailleurs, nous tenons à signaler que les étudiants ont particulièrement apprécié cette approche qui est également exploitée par ma collègue Minerva Rojas notamment lors d'autres cours portant sur l'analyse des manuels et sur l'évaluation du Master FLE d'Université Côte d'Azur. Cette approche sera développée grâce à la nouvelle maquette du Master qui intègre des cours sur la constitution et l'exploitation de corpus numériques. L'un des objectifs de notre équipe pédagogique est de sensibiliser les futurs enseignants à l'analyse quantitative des données textuelles comme support à l'analyse qualitative du français, que ce soit pour étudier les supports pédagogiques utilisables en cours ou pour évaluer les productions des apprenants de FLE. Notre philosophie vise à rendre rigoureux le processus d'enseignement-apprentissage du FLE, et à préparer nos étudiants à la recherche afin d'enrichir les démarches pédagogiques.

Références bibliographiques

Beacco, J.C. Blin, B. Houles, E. Lepage, S. Riba, P. (dir.). (2011). *Niveau B1 pour le français. Un référentiel*. Paris, Didier.

Beacco, J.C. Bouquet, S. Porquier, R. (dir.). (2004a). *Niveau B2 pour le français. Un référentiel*. Paris, Didier.

Beacco, J.C. Bouquet, S. Porquier, R. (dir.). (2004b). *Niveau B2 pour le français. Textes et références*. Paris, Didier.

Beacco, J.C. Lepage, S. Porquier, R. Riba, P. (dir.). (2008). *Niveau A2 pour le français. Un référentiel*. Paris, Didier.

Beacco, J.C. Porquier, R. (dir.). (2007). *Niveau A1 pour le français. Un référentiel*. Paris, Didier.

Boulton, A. Tyne, H. (2014). *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues* Paris, Didier.

Cavalla, C. (2019). Comment former les étudiants de Master FLE à l'utilisation pédagogique des corpus numériques ?. in Goes, J. Meneses-Lerin, L. Mangiante, J.M. Olmo F. Pineira-Tresmontant, C. *Apports et limites des corpus numériques en analyse de discours et didactique des langues de spécialité.* Editura Universitaria, 79-92, 978-606-14-1550-2. hal-02534091.

Cavalla, C. Loiseau, M. (2013). Scientext comme corpus pour l'enseignement ». in Tutin, A. Grossman, F. (eds.), *L'écrit scientifique : du lexique au discours. Autour de Scientext*, Rennes, PUR, 163-182.

Conseil de l'Europe. (2001). *Cadre Européen Commun de Référence pour les Langues.* Paris, Didier.

Conseil de l'Europe. (2018). *Cadre Européen Commun de Référence pour les Langues : volume complémentaire avec des nouveaux descripteurs*. https://rm.coe.int/cecr-volume-complementaire-avec-de-nouveaux-descripteurs/16807875d5

Fligelstone, S. (1993). Some reflections on the question of teaching, from a corpus linguistics perspective. *ICAME journal*, n°17, 87-109.

Holec, H. (1990). Qu'est-ce qu'apprendre à apprendre, *Mélanges pédagogiques*, 20, 75-87.

Johns, T. (1988). Implications et applications des logiciels de concordance dans la salle de classe. *Les langues modernes*, 82/5, 29-45.

Johns, T. King, P. (eds.). (1991). Classroom Concordancing. *English Language Research Journal*, 4, 47-61.

Kubler, N. (2014). Mettre en œuvre la linguistique de corpus à l'université. Vers une compétence utile pour l'enseignement/apprentissage des langues ?. [en ligne], 11-1 | 2014, mis en ligne le 07 janvier 2014. URL : http://journals.opendition.org/rdlc/1685

Renouf, A. (1997). Teaching corpus linguistics to teachers of English. in WICHMANN, A. FLIGELSTONE, S. McENERY, T. KNOWLES, G. (dir.), *Teaching and language corpora*, Harlow, Addison Wesley Longman, 255-266.

Riba, P. (dir.). (2016). *Niveau C1/C2 pour le français. Eléments pour un référentiel*. Paris, Didier.

Ruggia, S. Vanni, L. (2021). DeepFLE : la plateforme pour évaluer le niveau d'un texte selon le CECRL. *Dialogues et Cultures*, 66, 235-254.

Ruggia, S. (2021). La lecture contrôlée et assistée par l'analyse statistique des données textuelles : comment et pourquoi interroger un corpus numérique ?. *Le français dans le monde, Recherches et Applications,* « Langues et pratiques numériques : nouveaux repères et nouvelles littératies en didactique des langues ? », janvier, 69, 84-100.

Ruggia, S. Gaillat, T. (dir.). (2023). *Corpus* 24, « Les corpus numériques pour la didactique des langues : de la formation des enseignants à l'élaboration de systèmes automatiques », https://journals.openedition.org/corpus/7438.

Tagliante, C. (2005). *L'évaluation et le Cadre européen commun*, Paris, CLE International.

Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary, *System*, 21/3, 371-380.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins Publishing Company.

Penso dunque sono…convinto ! Pour une analyse quantitative des verbes d'opinion en français et en italien

Linda Sanvido
Institut des sciences du langage, Université de Neuchâtel
linda.sanvido@unine.ch

Introduction

Nous nous intéressons à des verbes qui appartiennent à la catégorie des *Source Introducing Predicates* (Saurí & Pustejovsky, 2009), désormais SIPs. Ces prédicats jouent un rôle fondamental pour l'interprétation de la proposition qu'ils accompagnent, puisqu'ils sont utilisés pour introduire une source supplémentaire dans le discours et que cette source représente un évaluateur de factualité de *p* (Saurí et Pustejovsky, 2009 : 236). Cette contribution se concentre spécifiquement sur des verbes français utilisés à la première personne du singulier du présent de l'indicatif, comme dans les exemples suivants :

a) **Je pense qu'**on mange très bien dans ce restaurant. Allons-y !
b) **Je trouve qu'**on mange très bien dans ce restaurant. Allons-y !

Ces deux SIPs sont utilisés pour exprimer le point de vue du locuteur, qui avance des réserves vis-à-vis du contenu de *p*. La fiabilité de *p* – et par conséquent, celle du locuteur – est plus affirmée dans b) que dans a). Ceci est dû au choix du verbe : *trouver* exprime un jugement qui est basé sur une expérience (cf. les analyses de Ducrot, 1980 et de Gosselin, 2015), tandis qu'avec *penser* le locuteur avance une simple hypothèse sur la vérité de *p*. On sera alors amené à suivre plus facilement la proposition de b) que celle de a).

Des verbes tels que *penser*, *croire*, *trouver*, *considérer*, *imaginer* mettent tous la subjectivité du locuteur au premier plan, en modalisant l'énoncé qu'ils introduisent, et cette modalisation fait partie de la réalisation de l'acte du locuteur. Nous essayerons de comprendre les effets rhétoriques et argumentatifs qui sont en jeu dans ce type de constructions et ce qui justifie le choix d'un SIP plutôt qu'un autre. Dans un deuxième temps, une analyse comparative avec l'italien sera menée.

Corpus et méthodologie

Afin de délimiter plus précisément le paradigme de verbes qui nous intéresse, nous utilisons une méthodologie *corpus-based* en faisant appel au logiciel TXM (Heiden et al., 2010). Le choix d'utilisation de corpus écrits[103] à caractère argumentatif plus ou moins marqué s'explique par les fréquences des SIPs, qui sont plus élevées dans ce type de texte que dans

[103] Le corpus PRES-FR est une version adaptée du corpus "Discours présidentiels français de 1958 à aujourd'hui", téléchargeable de Hyperbase Web (http://hyperbase.unice.fr/hyperbase/). Tous les autres corpus ont été constitués par nous-mêmes ou des membres de notre équipe de travail.

des textes plus normés et objectifs comme peuvent l'être des discours informatifs ou académiques (cf. Pamuksaç & Sanvido, soumis).

	Discours présidentiels français (PRES-FR)	**Le Monde (LM)**	**Discours présidentiels italiens (PRES-IT)**	**La Repubblica (REP)**
tokens	2'843'577	17'895'009	2'639'007	26'418'073
année	1958-2020	2010	1955-2021	2010-2015
nombre de textes	733 discours écrits	37'030 articles	2'194 discours écrits	52'996 articles
type de discours	politique	journalistique	politique	journalistique
langue	français	français	italien	italien

table 12. : **Corpus utilisés pour le français et l'italien.**

Pour établir notre paradigme de SIPs français, nous extrayons tous les verbes au présent de l'indicatif qui co-occurrent de façon statistiquement significative[104] à *je* et à *que* et nous gardons uniquement ceux qui apparaissent systématiquement dans nos deux corpus. Cette démarche permet de se focaliser sur les verbes les plus fréquemment utilisés dans ce type de configurations. Parmi ceux-ci, nous garderons uniquement ceux qui expriment une opinion par rapport à *p*, ou un jugement épistémique concernant la vérité de *p*. Ces verbes peuvent être remplacés par *à mon avis* sans que le sens de *p* change :

a) **Je crois que/ A mon avis** le temps va changer.
b) **J'imagine que/ A mon avis** tu es fatiguée.

Des verbes de sentiment tels que *se réjouir*, *aimer,* de volition tels que *souhaiter*, *espérer* ou des verbes factifs comme *savoir, observer* seront exclus. En effet, ils ne passent pas le test de *à mon avis* car ils apportent une contribution sémantique supplémentaire à *p* :

c) **J'espère que/ ??A mon avis** je pourrai partir à la mer cet été.
d) **J'observe que/ ??A mon avis** les prix montent rapidement.

Ces deux SIPs ne véhiculent aucun jugement quant à la validation du contenu de *p*, ils expriment une activité psychocognitive portant sur le contenu de *p* : un espoir pour c) et une constatation pour d). Rentrerons alors dans notre paradigme uniquement les SIPs compatibles avec le test de remplacement : il s'agit de verbes traditionnellement classés comme prédicats de croyance (*penser*, *croire,* …), d'opinion (*trouver*, *considérer,* …), mais aussi de communication (*dire*).

Ce paradigme sera interrogé par le biais d'une analyse quantitative exploitant les scores de cooccurrence pour analyser l'environnement lexical des SIPs. Cette démarche s'explique par le principe bien connu qu'un mot est mieux compris s'il est étudié avec les mots qui l'accompagnent (Harris 1954 ; Firth 1957 ; Sinclair 1996 ; Tognini-Bonelli 2001 ; Rundell 2018). Puisque les SIPs jouent un rôle important dans l'implication de la subjectivité du locuteur, une attention particulière sera portée aux cooccurrences spécifiques qui sont également en jeu dans le processus d'expression d'une attitude sur un état de choses. Nous

[104] TXM propose de mesurer la cooccurrence significative entre deux formes avec l'indice de cooccurrence de Lafon (1981). Le calcul de cet indice prend en compte quatre éléments : la fréquence d'une forme X au côté d'une forme Y dans une fenêtre restreinte prédéfinie, la fréquence totale de X et de Y dans le corpus, la taille de l'empan dans lequel Y apparait et la taille totale du corpus. L'indice seuil, qui établit que la co-occurrence entre deux mots ou chaines de mots est statistiquement significative, est fixé à 2. Plus élevé est l'indice entre deux mots, plus forte en résulte l'attraction.

nous concentrons en particulier sur des marqueurs modaux (verbes exprimant une possibilité et une nécessité) et des marqueurs axiologiques (adjectifs et adverbes subjectifs, cf. Kerbrat-Orecchioni 1980 : 83, 118).

Résultats

Nous avons commencé par questionner, dans les deux corpus français, l'empan droit (10 items) de nos SIPs, afin d'étudier le contenu de la complétive qu'ils introduisent. Ces premières recherches permettent déjà de dresser une description sémantique de certains de nos verbes.

- *je pense que* et *je crois que* sont les SIPs le plus sensibles aux catégories subjectives, puisqu'ils attirent systématiquement des marqueurs épistémiques, déontiques, mais aussi des adjectifs axiologiques ;
- *je trouve que* est sensible, comme *je pense que* et *je crois que*, aux marqueurs axiologiques. Cependant, il se distingue de ces deux autres SIPs par son imperméabilité aux catégories des marqueurs modaux (nécessité et possibilité) ;
- *je dis que* et *je considère que* se combinent de manière statistiquement significative avec des marqueurs axiologiques et des modalités déontiques uniquement (respectivement avec les verbes *falloir* et *devoir*).

Grâce à ces premières observations, nous avançons quelques hypothèses sur le comportement sémantico-pragmatique de nos SIPs : *je trouve que* n'aide pas à prendre des précautions vis-à-vis du contenu de *p* – comme peuvent le faire *je crois/pense que* – mais permet de mettre le focus sur l'opinion du locuteur en question, en la contrastant aux idées d'autrui. *Je considère/dis que* renforcent les propos du locuteur en travaillant sur la force illocutoire associée à *p* (dans les exemples tirés de nos corpus, l'illocution se traduit souvent par un conseil ou une requête). Afin d'avoir un regard comparatif entre langues apparentées, ces résultats seront mis en perspective avec les formes correspondantes en italien. Grâce à cette approche comparative, il sera possible non seulement de vérifier si les deux paradigmes ont les mêmes propriétés, mais aussi de mettre en évidence les différences de modalisation, ou de pragmaticalisation, de ces formes dans deux langues romanes.

Références bibliographiques

Ducrot, O. et al. (1980). *Les mots du discours*. Paris : Minuit.

Firth, J.R. (1957). *Papers in linguistics 1934–51*. Oxford University Press.

Gosselin, L. (2015). L'expression de l'opinion personnelle : « Je crois/pense/trouve/considère/estime que p ». *L'information grammaticale*, 144, 34-40.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146-162.

Heiden S, Magué J.-P., and Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*. Rome, Italie. URL : http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf

Kerbrat-Orecchioni, C. (1980). *L'énonciation. De la subjectivité dans le langage*. Paris : Armand Colin.

Lafon P. (1981). Analyse lexicométrique et recherche des cooccurrences. *MOTS*, 3, 95-148.
Pamuksaç, A. & Sanvido, L. (soumis). *Distinguer les genres à travers l'étude des traces du locuteur : analyse sur corpus et application en FLE*. Sixième congrès international franco-espagnol E-GRAPHELES, Universitat Politècnica de Valènca, Valence (Espagne), 19-21 octobre 2022.
Rundell, M. (2018). Searching for extended units of meaning-and what to do when you find them. *Lexicography*, 5(1), 5-21.
Saurí, R., & Pustejovsky, J. (2009). FactBank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3), 227-268.
Sinclair, J. (1996). The search for units of meaning. *Textus*, 9(1), 75-106.
Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: J. Benjamins.

Sentence Processing in Translation: A Corpus Approach

Maya Sfeir [1] et Georgeta Cislaru [2]
[1] Department of English, American University of Beirut
[2] Laboratoire MoDyCo, Université Paris Nanterre
ms213@aub.edu.lb, gcislaru@parisnanterre.fr

Introduction

Translation, like writing, is one type of text production; both have the purpose of producing a coherent text for a specific audience following a process of planning, drafting, and revision (Dam-Jensen & Heine, 2013; Dragsted & Carl, 2013). However, unlike writing, which indirectly builds on other sources and texts, translation necessitates an existing source text (Dam-Jensen & Heine, 2013; Dragsted & Carl, 2013), and subsequently, the reading and reformulation of already existing linguistic units. The translation process is itself influenced by a multitude of cognitive, linguistic, and functional variables, such as the knowledge of the source and target languages and cultures (Hayes, 2012; Neubert, 2000), and translator expertise and style, and even typing speed (Dam-Jensen & Heine, 2013).

Integral to both the writing and translation processes is the production of bursts, defined as spontaneously produced units of text occurring within a two-second pause threshold, and shaped by cognitive and functional considerations (Chenoweth & Hayes 2001; Hayes, 2012; for a linguistics insight see Cislaru & Olive, 2017, 2018). Previous research focusing on the production of bursts during the writing process has examined their grammatical structure (Cislaru & Olive 2018), their length in relation to writer expertise (Kaufer et al. 1986), their role in text organization and segmentation (Cislaru & Olive, 2018; Feltgen et al., 2022), the syntactic deconstruction potentially underlied by bursts (Cislaru & Olive, 2021), and the behavior of subject clitics during the dynamics of writing (Feltgen et al., 2023).

In translation, writing bursts have mostly been examined using talk-aloud protocols (Gerloff, 1986, 1988; Kussmaul & Tirkkonen-Condit, 1995). Results have revealed that while students, bilinguals, and professional translators all work with “small, syntactic units”, bilinguals and professionals could process larger chunks of discourse, and they revised and edited their translation more than students (Gerloff, 1988, p. 148). Other studies focused on strategies that might lead to successful or unsuccessful translations (Gerloff, 1986), highlighting the fact that non-professionals approach translation as a linguistic task (Tirkkonen-Condit, 1990).

However, given that these studies have mostly relied on talk-aloud protocols, they have not treated bursts produced during translation writing in real time, and as a result, little is known about the linguistic and syntactic structure and textual functions of these translation writing bursts. Recent advances in corpus and computational tools and approaches, in particular keystroke logging programs and annotated corpora of keystroke logs, make it possible today to examine the production of writing bursts during translation in real time (Farnoud, 2014). Our study, exploratory in nature, seeks to examine, grammatically and semantically, writing bursts produced during the translation process, with a focus on sentence processing.

Corpus et méthodologie

Corpus

We analyze writing process data delivered by Inputlog, a keystroke logging software (Leijten & Van Waes 2013). Our corpus consists of 19 translations of medical texts from English to French produced in French by University of Montréal second- and third-year undergraduate translation students, and 19 control texts produced in French by the same students. The corpus includes both the final texts and the writing bursts produced by the students. It has been partially annotated for morpho-syntactic categories and dependency syntactic structures. Our annotated corpus of keystroke logs thus makes it possible to compare sentence processing during both writing and translation activities, as students worked towards producing different genres, essays and translations.

Méthodologie

Our analysis will examine bursts of writing in the light of syntactic frontiers, word order, and idiomatic structures, in order to grasp i) their nature and ii) their evolution into sentences, all while taking into consideration type of writing activity constraints and genres. The nature of our corpus will enable us to compare keystroke logs generated during the writing and translation of medical texts, and thus sentence processing resulting from the two activities. Taking into consideration the translators' expertise and grammatical differences between French and English (Mailhard, 2000), our study and findings pave the way for a better understanding of translation competence (Schäfner & Adab, 2000), with implications for the teaching and learning of translation and the training of translators.

Résultats

Empirical research results will be shared and discussed during the presentation. The analysis will take into consideration sentence processing in the translated texts as well as the written productions (the control texts) to better account for the particularities pertaining to sentence processing in translation.

Références bibliographiques

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: generating text in L1 and L2. *Written Communication 18*(1), 80-98. https://doi.org/10.1177/0741088301018001

Cislaru, G., & Olive, T. (2017). Segments répétés, jets textuels et autres routines. Quel niveau de pré-construction?. *Corpus*, *17*, 1-21. https://doi.org/10.4000/corpus.2846

Cislaru, G., & Olive, T. (2018). Bursts of written language as performance units for the description of genre routines. In D. Legallois, T. Charnois, & M. Larjavaara (Eds.), *The grammar of genres and styles: From discrete to non-discrete units* (pp. 219-246). De Gruyter.

Cislaru, G., & Olive, T. (2021). Que peut nous apprendre l'écriture enregistrée en temps réel au sujet des figures de construction?. *L'Information grammaticale 169,* 21-29.

Dam-Jensen, H., & Heine, C. (2013). Writing and translation process research: Bridging the gap (Introduction). *Journal of Writing Research*, *5*(1), 89-101. https://doi.org/10.17239/jowr-2013.05.01.4

Dragsted, B., & Carl, M. (2013). Towards a classification of translator profiles based on eye-tracking and keylogging data. *Journal of Writing Research*, *5*(1), 133-158. https://doi.org/10.17239/jowr-2013.05.01.6

Farnoud, E. (2014). Processus de la traduction: Charge cognitive du traducteur. *Corela*, *12*(2), 1-17. Doi: 10.4000/corela.3615

Feltgen, Q., Cislaru, G., & Benzitoun, C. (2022). Étude linguistique et statistique des unités de performance écrite: le cas de et. In *SHS Web of Conferences, 138*, 1-17. https://doi.org/10.1051/shsconf/202213810001

Feltgen, Q., Lefeuvre, F., & Legallois, D. (2023). Sujet clitique et dynamique de l'écrit: un éclairage par les jets textuels. *Discours, 32.*

Gerloff, P. (1986). Second language learner's reports on the interpretive process: Talk-aloud protocols of translation. In J. House & S. Blum-Kulka (Eds), *Interlingual and intercultural communication* (pp. 243-262). Tübingen Narr.

Gerloff, P. A. (1988). *From French to English: A look at the translation process in students, bilinguals, and professional translators* (Order No. 8823316). Available from ProQuest Dissertations & Theses Global. (303679401). https://www.proquest.com/dissertations-theses/french-english-look-at-translation-process/docview/303679401/se-2

Hayes, J. R. (2012). Evidence from language bursts, revision, and transcription for translation and its relation to other writing processes. In M. Fayol, D. Alamargot, & V. W. Berninger (Eds.), *Translation of thought to written text while composing: Advancing theory, knowledge, research methods, tools, and applications* (pp. 15–25). Psychology Press.

Kaufer, D. S., Hayes, J. R., & Flower, L. (1986). Composing written sentences. *Research in the Teaching of English*, 121-140. http://www.jstor.org/stable/40171073

Kussmaul, P. & Tirkkonen-Condit, S. (1995). Think-Aloud Protocol Analysis in Translation Studies. *TTR*, *8*(1), 177–199. https://doi.org/10.7202/037201ar

Mailhac, J-P. (2000). Levels of speech and grammar when translating between English and French. In C. Schäffner & B. Adab (Eds.), *Developing translation competence* (pp. 33-50). John Benjamins.

Neubert, A. (2000). Competence in language, in languages, and in translation. In C. Schäffner & B. Adab (Eds.), *Developing translation competence* (pp. 3-18). John Benjamins.

Schaffner, C., & Adab, B. Developing translation competence: Introduction. In C. Schäffner & B. Adab (Eds.), *Developing translation competence* (pp. vii-xvi). John Benjamins.

Tirkkonen-Condit, S. (1990). Professional vs. non-professional translation: A think-aloud protocol study. *Learning, keeping and using language* 2, 381-394.

Rendre un grand corpus oral accessible pour la didactique du FLE : le projet ESLOFLEU

Marie Skrovec [1], Chloé Tahar [1], Flora Badin [1], Britta Thörle [2]
[1] Laboratoire Ligérien de Linguistique, Université d'Orléans
[2] Département de Romanistik, Université de Siegen
marie.skrovec@univ-orleans.fr, chloe.tahar@univ-orleans.fr, flora.badin@univ-orleans.fr,
thoerle@romanistik.uni-siegen.de

Introduction

Le corpus ESLO (Enquêtes SocioLinguistiques à Orléans), l'un des plus grands corpus oraux disponibles pour le français avec ses 600h d'enregistrements audio transcrits (422h en libre accès), constitue un important terrain d'observation du français parlé contemporain pour les linguistes. Structuré selon plusieurs axes, ce « portrait sonore » d'Orléans par ses habitants permet d'analyser le français parlé dans les années 60 et aujourd'hui, en accédant à une diversité de locuteurs dans des situations de communication variées. Le projet ESLO-FLEU (ESLO pour le Fle et la Linguistique dans l'Enseignement Universitaire), qui repose sur une collaboration entre le LLL à l'université d'Orléans et le département des langues romanes à l'université de Siegen, vise à rendre le corpus ESLO accessible à une communauté plus large d'étudiants et d'enseignants dans le domaine du FLE universitaire. Nous montrons comment nous avons sélectionné et structuré des données issues de ce « réservoir », dans une approche de prédidactisation. Nous nous concentrons ici sur l'étape de sélection et d'indexation d'extraits, préalable à l'élaboration d'une ressource numérique. Cette ressource, constituée de plusieurs modules correspondant aux objectifs de formation des spécialistes du français en contexte universitaire à l'étranger (départements de français et d'études françaises et/ou romanes), est testée en 2022-2023 dans le cadre de cours de spécialité en sciences du langage, linguistique sur corpus et civilisation à l'université de Siegen. Nous expliquons également les choix techniques liés à un double objectif, celui de veiller à la réutilisabilité de ce travail d'enrichissement des données en linguistique (cumulativité), tout en garantissant l'accessibilité des données pour les non linguistes.

État de l'art

La question de l'utilisation des corpus des linguistes pour l'enseignement des langues étrangères est de grande actualité. Depuis plus d'une dizaine d'années, la méthode d'apprentissage sur corpus a été expérimentée, notamment pour l'anglais, domaine précurseur en la matière, à partir de corpus écrits ou oraux (Sinclair 2004, Boulton & Tyne 2014). Parallèlement, on note, à la suite de l'intérêt porté en linguistique pour le français parlé (Blanche-Benveniste 2010, Blanche-Benveniste & Martin 2010) un intérêt en didactique du FLE pour l'oralité et les spécificités du français parlé contemporain (Weber 2013). Parmi les linguistes, nombreux sont ceux à collaborer avec des didacticiens en vue de proposer des ressources pour la didactique du FLE (cf. les différents projets décrits dans Detey et al. 2011, André 2016, Etienne & Jouin 2019, Surcouf & Ausoni 2018).

Corpus et méthodologie

Corpus

Le corpus ESLO présente la particularité d'avoir été constitué en 2 temps (ESLO1 1968-1971, ESLO2 depuis 2008) selon une méthodologie similaire. Le projet trouve son origine à la fin des années 1960 lorsqu'une équipe franco-britannique, composée notamment d'enseignants de français, se fixe pour objectif de créer un corpus afin de constituer un ensemble de documents authentiques pour l'enseignement du français qui aboutit à la publication d'un manuel, *Les Orléanais ont la parole* (Biggs et Dalwood, 1976). L'équipe a également des objectifs de diffusion et d'analyse sociolinguistique (Blanc et Biggs, 1971) et réalise ainsi le « portrait sonore d'une ville » en tenant compte de « l'identité sociale de chaque locuteur et de la situation de communication dans chaque cas » (Lonergan et al., 1974 : 4). Depuis 2008, le laboratoire CORAL devenu LLL a entrepris de constituer un 2e volet, sous forme d'un échantillon comparable à ESLO1 basé sur la situation de l'entretien sociolinguistique, qui permet d'accéder aux représentations des Orléanais sur leur ville, la langue, leur rapport à la norme, etc. En outre, l'accent est mis dans ESLO2 sur la diversité des situations dans un contexte urbain, de manière à « répondre à un objectif de couverture étendue de la diversité des pratiques linguistiques dans la vie quotidienne » (Baude 2015 : 124). Parallèlement, un travail sur l'élaboration d'une typologie des situations de communication a été mené (Baude & Guerin 2014, Baude 2015), basée sur l'idée de continuum variationnel proposée par les romanistes allemands Söll (1974) puis Koch & Oesterreicher (2001) et relayée en France par Gadet (2003). Si plusieurs chercheurs du LLL s'intéressent à la question des transferts didactiques de l'accessibilité du corpus ESLO pour l'enseignement (Skrovec 2019, 2020 ; Hamma 2019), force est de constater qu'en l'absence d'un travail de pré-didactisation de ces données, la ressource reste peu accessible aux non spécialistes en linguistique de corpus en raison de la taille du corpus et de la technicité des outils d'exploration.

Méthodologie : sélection et indexation thématique

L'exploration du corpus a été guidée par la recherche d'extraits présentant un intérêt à plusieurs niveaux. Il s'agissait de :

- valoriser la dimension patrimoniale du corpus en sélectionnant des sujets de conversation saillants, des passages représentatifs du grand corpus ;
- rendre compte de la diversité des locuteurs et des situations d'interaction représentées dans le corpus ;
- identifier des extraits utilisables en contexte pédagogique (FLE universitaire en premier lieu, mais aussi FLE généraliste, voire dans d'autres disciplines), de par leur longueur et contenu. Les frontières ont été définies de manière à circonscrire des séquences d'une durée de 2 à 4 minutes maximum présentant une cohérence thématique et conversationnelle.

Le travail d'exploration-sélection a permis de faire émerger des regroupements thématiques, selon un processus prévoyant plusieurs étapes : recherche de tout type d'extrait d'intérêt, constitution de collections par regroupement thématique (recherche de similarités thématiques entre extraits), suppression d'extraits au thème marginal. L'intégralité des 150 extraits a fait l'objet d'une double indexation au minimum : chaque extrait appartient à un grand ensemble thématique (8 valeurs), et reçoit une sous-spécification thématique (43 valeurs). Parmi ces extraits, 76 intègrent un des 5 modules de formation (livrables du projet) et reçoivent une étiquette supplémentaire (5 valeurs).

L'indexation, pensée pour un public non-expert, a été réalisée en privilégiant le caractère intuitif des étiquettes. Elle vise d'une part (i) à identifier des catégories culturelles de sens commun, qui reflètent plusieurs aspects de la vie quotidienne des Orléanais (Métiers, Loisirs, Société, Vivre à Orléans) ; d'autre part (ii) à identifier des thèmes saillants émergeant dans certains dynamiques conversationnelles (Récits, Conversations, Savoirs, Témoignages). Par exemple, la sous-spécification thématique de l'ensemble des Récits rend compte du fait que les récits conversationnels sont le locus privilégié de l'élaboration de sujets conversationnels prototypiques (Récits familiaux, Expériences personnelles, Premiers emplois, Projets d'avenir, Remémorations d'histoires). La pertinence de l'annotation thématique a été évaluée en fonction de l'accord intersubjectif des deux chercheuses impliquées dans ce travail.

Méthodologie : choix des outils d'annotation et de consultation La ressource est élaborée sous deux formes :

- Une version TXM (Heiden 2010) qui permet le balisage et l'annotation dans l'intégralité du corpus (correspondant à la partie accessible en ligne), pour l'utilisateur linguiste expert formé à l'annotation et l'interrogation avancée.
- Une version HTML sous forme de fichiers consultables sur un navigateur, pour une utilisation facilitée sans système de requête : consultation des extraits alignés au son grâce à une barre de lecture, ainsi qu'aux annotations, pour l'utilisateur linguiste non formé à l'annotation, l'apprenti linguiste ou le non linguiste.

Résultats

L'exploration a fait émerger 8 thématiques, répertoriées ci-après (figure 1).

Thème	Nombre d'extraits	Pourcentage
Conversations	15	10%
Loisirs	21	14%
Métiers	12	9%
Récits	24	15%
Savoirs	14	9%
Société	17	11%
Témoignages	24	16%
Vivre à Orléans	23	15%
Total	150	100%

table 13. : figure . 1 Regroupement des extraits par thème

Les extraits sont également indexés selon un système de sous-spécification thématique (figure 2).

Thèmes	Sous-thèmes	Nombre d'extraits
Vivre à Orléans		**24**
	Équipements collectifs	4
	Fêtes orléanaises	5
	Les Orléanais	3
	Logement	4
	Modes de vie	5
	Sociolecte	3

table 14. : figure . 2 Exemple des sous-spécifications thématiques du thème « Vivre à Orléans »

Après ce premier travail, une deuxième sélection de 76 extraits parmi les 150 de la sélection initiale a été effectuée en vue de constituer 5 modules pour la formation universitaire dans le cadre de cours de spécialité (linguistique interactionnelle, linguistique de corpus, sociolinguistique, didactique du FLE, civilisation). Visant une meilleure connaissance du français tel qu'il se parle, trois de ces modules permettent l'observation, la conceptualisation et la systématisation d'objets linguistiques caractérisant les interactions orales en français : « Module 1 : D'une situation à l'autre » (14 extraits ; variation diaphasique, liaison et constructions interrogatives), « Module 2 : Le parler jeune » (18 extraits ; variétés jeunes et usages des marqueurs discursifs, lexique familier, dispositifs syntaxiques spécifiques : dislocations), « Module 3 : Récits en interaction » (13 ; émergence des récits en interaction, discours rapporté et dispositifs syntaxiques spécifiques : clivées et présentatives complexes). Deux autres modules, à finalité socio-culturelle, « Voix variées » (11 extraits) et « Parole d'Orléanais » (20 extraits), rendent hommage au travail réalisé par les pionniers du corpus ESLO. « Parole d'Orléanais » allie thématiques contemporaines (fêtes de Jeanne d'Arc, sociolecte orléanais, centre-ville et quartiers, perception des Orléanais par eux-mêmes) et sujets historiques (mai 1968 et lutte des classes, Trente Glorieuses, pieds noirs d'Algérie), constituant une collection intéressante pour une exploitation dans un cours à thématique culturelle (Civilisation française, Culture et société, etc.).

Conclusion

Le projet ESLO-FLEU s'inscrit dans une démarche de transfert des recherches en linguistique et d'échanges avec le domaine de la didactique, au niveau universitaire et au-delà. Il s'agit ainsi de valoriser les corpus oraux au-delà de la communauté des linguistes, pour répondre à des orientations formulées en didactique des langues aujourd'hui (importance des compétences interactionnelles, exposition des apprenants à des conversation spontanées en contexte naturel). Prévus pour la formation universitaire des futurs enseignants de FLE dans le cadre de ce projet, les modules visant des objectifs linguistiques précis reposent sur un travail d'annotation des données issu des méthodes actuelles en linguistique. Mais cette ressource, conçue également pour un usage non expert, peut être utilisée par des non linguistes : dans la mesure où elle constitue une collection d'extraits sonores illustrant les usages du français parlé hier et aujourd'hui à partir de contenus socio-culturels divers, elle est susceptible d'intéresser les enseignants de FLE.

Références bibliographiques

André Virginie, 2016, FLEURON : Français Langue Étrangère Universitaire – Ressources et Outils Numériques. Origine, démarches et perspectives. *Mélanges Crapel 37*, p.69-92.

Baude, Olivier, 2015, *Observatologie : Vers une science de l'adéquation observationnelle en linguistique*. Linguistique. Thèse d'habilitation, Université Paris Ouest Nanterre la défense.

Baude, Olivier & Guerin, Emmanuelle, 2014, Pourquoi et comment dresser le portrait sonore d'une "grande ville" ? L'exemple d'ESLO2. *Les métropoles francophones en temps de globalisation*, Juin 2014, Nanterre, France.

Biggs, Patricia & Dalwood, Mary, 1976, *Les Orléanais ont la parole*, Londres : Longman.

Blanc, Michel & Biggs, Patricia, 1971, L'enquête socio-linguistique sur le français parlé à Orléans. *Le français dans le monde* 85 : 16-25.

Blanche-Benveniste, Claire, 2010, *Approches de la langue parlée en français*. Paris : Ophrys.

Blanche-Benveniste, Claire & Martin, Philippe, 2010, *Le français. Usages de la langue parlée*. Leuven/Paris : Peeters.

Boulton, Alex & Tyne, Henry, 2014, Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues, Paris : Didier.

Detey, Sylvain & Durand, Jacques & Laks, Bernard & Lyche, Chantal, 2011, *Les variétés du français parlé dans l'espace francophone. Ressources pour l'enseignement*. Paris : Ophrys.

Etienne Carole & Jouin Emilie, 2019, Constituer des ressources pédagogiques pour enseigner le français oral à partir des recherches menées en interaction. In Gajo L., Luscher J.-M., Racine I., Zay F. (eds), *Variation, plurilinguisme et évaluation en français langue étrangère*. Bern : Peter Lang, p.225-240.

Gadet, Françoise, 2003, *La variation sociale en français*. Paris : Ophrys.

Hamma, Badreddine, 2019, Quand l'interaction n'est pas là, la souris est mangée par le chat ! Remarques sur l'enseignement du passif en classe de français. In : *Linguistique interactionnelle, grammaire de l'oral et didactique du français.* Anne-Sophie Calinon, Badreddine Hamma, Katja Ploog et Marie Skrovec (éds.). Besançon : Presses Universitaires de Franche-Comté, 237-262.

Heiden, Serge & al., 2010, TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement, in *Actes du 10th International Conference on the Statistical Analysis of Textual Data*, Rome, Edizioni Universitarie di Lettere Economia Diritto 2, 1021-1032.

Koch, Peter & Oesterreicher, Wulf, 2001, Gesprochene Sprache und geschriebene Sprache / Langage parlé et langage écrit. In: G. Holtus, M. Metzeltin, Ch. Schmitt (éds), *Lexikon der Romanistischen Linguistik*, Bd. I/2, 584-627.

Lonergan J., Kay J. & Ross J. 1974, Étude sociolinguistique sur Orléans, catalogue des enregistrements. Colchester : Multigraphié.

Sinclair, John, 2004, *How to use corpora in language teaching*. Amsterdam/Philadelphia : John Benjamins.

Skrovec, Marie, 2019, « Grammaire et corpus oraux : réflexions pour un prolongement didactique de travaux réalisés sur ESLO », in : Calinon, A.-S., Hamma, B., Ploog, K. & Skrovec, M. (éds.) : *Linguistique interactionnelle, grammaire de l'oral et didactique du français*, actes des Journées d'étude Linguistique et Didactique, Orléans, 23 mai et Besançon, 5 décembre 2014, Besançon : Presses Universitaires de Franche-Comté, 263294.

Skrovec, Marie, 2020, « Le futur en français parlé, simple ou périphrastique ? Pour une adaptation des corpus oraux en didactique du FLE ». In : La didactisation du français vernaculaire, *Syntaxe et Sémantique*, 113-151.

Söll, Ludwig, 1974, *Gesprochenes und geschriebenes Französisch*. Berlin : E. Schmidt.

Surcouf Christian & Ausoni Alain, 2018, « Création d'un corpus de français parlé à des fins pédagogiques en FLE : la genèse du projet FLORALE », EDL (Études en didactique des langues), n°31, pp. 71-91.

Weber, Corinne, 2013, Pour une didactique de l'oralité. Enseigner le français tel qu'il est parlé, Paris : Didier.

Le corpus oral ESLO comme ressource didactique pour la formation universitaire en FLE et sciences du langage : l'exemple d'un module sur les *mots du discours*

Marie Skrovec [105], Britta Thörle [106], Chloé Tahar [1], Flora Badin[1]
[1] Laboratoire Ligérien de Linguistique, Université d'Orléans
[2] Département de Romanistik, Université de Siegen
marie.skrovec@univ-orleans.fr, thoerle@romanistik.uni-siegen.de, chloe.tahar@univ-orleans.fr, flora.badin@univ-orleans.fr

Introduction

Depuis une trentaine d'années, des chercheurs à l'interface entre la linguistique et la didactique des langues se penchent sur la question de savoir comment les corpus de langues peuvent être utilisés d'un point de vue didactique. Si les initiatives dans ce domaine se sont d'abord concentrées sur l'espace anglophone, on constate depuis une dizaine d'années une augmentation des projets concernant les corpus francophones. La présente contribution est issue du projet ESLO-FLEU (ESLO pour le FLE et la Linguistique dans l'Enseignement Universitaire), qui a pour objectif de développer des ressources d'enseignement-apprentissage basées sur des corpus pour le français langue étrangère dans le contexte universitaire. Issus du corpus ESLO (Enquêtes SocioLinguistiques à Orléans[1]), des modules d'enseignement-apprentissage sont élaborés sur différents thèmes à l'interface entre didactique des langues et sciences du langage et visent à développer chez les apprenants une meilleure connaissance du français tel qu'il se parle. Ces modules, qui ne se présentent pas comme des unités d'enseignement intégralement didactisés et scénarisés, sont conçus comme choix d'extraits de corpus illustrant une certaine thématique et facilitant, par son annotation, l'analyse d'un ou de plusieurs phénomènes linguistiques. Dans ce qui suit, le projet sera présenté à l'exemple du module « Le parler jeune » qui traite, entre autres, des mots du discours (MD) en français parlé[2]. Si les MD ont été reconnus comme un outil central de la compétence communicative et interactionnelle dans la didactique des langues étrangères, les concepts didactiques et le matériel pour leur enseignement font cependant défaut. Le travail avec des corpus de conversations naturelles, intégrées dans des situations d'interaction authentiques, représente ici une ressource particulièrement précieuse. Les objectifs de cette contribution sont 1) de présenter la méthodologie de la constitution et de l'annotation du

[105] http://eslo.huma-num.fr/

[106] Ce module est construit pour familiariser les apprenants avec un ensemble d'objets actuellement décrits en linguistique sous l'appellation « marqueurs discursifs ». Devant certaines difficultés de catégorisation liées à la nature-même de ces marqueurs (formes polysémiques, ensemble non homogène d'un point de vue syntaxique, etc.), et pour répondre à une nécessaire adaptation terminologique pour un public non expert, le terme "mot du discours", renvoyant aux *Gesprächswörter* de la tradition germanophone (cf. Koch/Oesterreicher 1990), a semblé pertinent pour désigner une diversité de phénomènes incluant entre autres interjections, connecteurs et marqueurs de l'interaction.

module en vue de sa finalité didactique, 2) de rapporter une expérimentation didactique sur la base du module et 3) de présenter les résultats pratiques de ce travail sous formes de « livrables ».

État de l'art : corpus oraux et enseignement des mots du discours

Les MD sont un élément central de la compétence discursive telle qu'elle est définie dans le CECRL (cf. ch. 5.2.3.1). Ce sont des moyens linguistiques indispensables lorsqu'il s'agit de commencer une conversation, de la maintenir ou de la terminer, de prendre ou de céder le tour de parole, d'introduire ou de conclure un sujet ou de s'assurer de la compréhension mutuelle. Alors que la pertinence des moyens linguistiques – et particulièrement des connecteurs – pour la cohérence et de la cohésion du texte est mise en évidence de manière relativement concrète dans le CECRL, le rôle des MD pour les autres aspects de la compétence discursive reste vague (cf. CECRL 5.2.3.1). Les études sur l'acquisition des MD en langue étrangère montrent toutefois que les apprenants disposent d'un répertoire extrêmement limité de MD, qui ne se différencie sur les plans formel et fonctionnel qu'à des niveaux très avancés (Borreguero Zuloaga/Thörle 2016). Les apprenants sont ainsi privés d'un prérequis important pour une participation compétente à la conversation. D'où la nécessité d'une prise en compte plus systématique des corpus de conversations naturelles en interaction authentique dans la didactique des langues, qui permettent l'observation des fonctions pragmatiques et sociales des MD et contribuent – contrairement aux purs inventaires de connecteurs – à une prise de conscience de leur fonctionnement (cf. aussi Beeching 2014). Dans le cadre de la formation de futurs enseignants de FLE, les MD sont un thème profitable en cours de linguistique française qui permet particulièrement bien d'étudier la relation entre langue et interaction.

Corpus et méthodologie

De précédentes études sur le corpus ESLO ont observé une fréquence importante de l'usage des MD dans les enregistrements impliquant des locuteurs jeunes, ceux-ci privilégiant par ailleurs leurs emplois les plus pragmaticalisés (voir Skrovec et al. 2022 pour *après*, Abouda 2022 pour *du coup*). C'est ce constat qui a guidé le choix de constituer un module de formation sur les MD à partir d'une sélection d'extraits sélectionnés selon l'âge des locuteurs.

Le module « Le parler jeune », composé de 20 extraits issus de situations informelles, implique majoritairement des locuteurs de la tranche d'âge 15/25 ans. Les interactions se caractérisent par une interactivité très élevée et une absence de préparation. Ce sous-corpus, de 10 000 mots environ et annoté sous TXM, a tout d'abord fait l'objet d'un repérage manuel, non exhaustif mais large, de formes susceptibles de recevoir une analyse linguistique comme marqueurs discursifs/mots du discours. On y dénombre 2125 occurrences pour 48 types de MD. Une annotation plus précise a ensuite été menée sur 23 MD pour spécifier ces formes d'un point de vue sémantico-pragmatique, selon 4 grands ensembles fonctionnels.

- Le domaine *Interaction* regroupe les marqueurs relatifs à la structure de l'interaction et à la co-construction des tours (*alors* en ouverture de tour, *hm* de réception), la gestion intersubjective (*mais* de désaccord) et à la gestion des connaissances partagées (*hein* de sollicitation de l'approbation, *voilà* de ratification) ;
- Le domaine *Formulation* comprend les MD associés au travail de formulation et de gestion de la progression discursive, au sens de Gülich (1993), comme les procédures

de correction (*enfin*, *quoi*), d'hésitation (*euh*), d'approximation (*genre*), d'atténuation (*quand même*), de proposition de formulation (*voilà*), de précision (*mais*), etc.

- Le domaine *Connexion* recense les emplois comme connecteurs logico-argumentatifs, pouvant porter sur du contenu propositionnel autant que sur des éléments contextuels explicites ou implicites, inférables de la situation ou des savoirs partagés, comme les emplois consécutifs de *donc* ou *du coup*, les emplois contrastifs de *mais* ou *quand même*, etc.
- Le domaine *Information* identifie les emplois liés à la gestion des topics conversationnels, comme les procédés d'activation de nouveau topic (*alors*), de développement topical (*du coup*), de parenthèse (*bon*), ou de clôture de topics (*voilà*, *bref*).

Ces annotations sont disponibles au format TXM pour les utilisateurs experts ou au format html pour une utilisation en classe, avec fonction afficher/masquer. Ainsi, il est possible de prévoir une phase de découverte du document sonore transcrit sans les annotations, puis d'afficher l'ensemble des MD repérés et leurs annotations. Cela doit permettre de guider les apprenants dans l'observation des formes et la formulation d'hypothèses sur leurs différentes fonctions pragmatiques. Un guide d'annotation, à destination des utilisateurs experts ou semi-experts, est en passe d'être finalisé et sera disponible avec l'intégralité du corpus annoté (version 2) sur Ortolang[107].

Application en cours de linguistique française à l'étranger

Un deuxième objectif du projet, outre la création de la ressource, est l'application et l'évaluation du module en contexte pédagogique. Un prototype du module "Parler jeune" a ainsi été utilisé à l'université de Siegen dans un séminaire de linguistique française pour les étudiants de licence avancés (niveau Bachelor en Allemagne). Dans ce cours, les participants travaillaient avec les extraits sonores transcrits, mais pas encore annotés, étant donné que l'annotation était encore en cours à l'époque. Le groupe cible du séminaire était en grande partie constitué d'étudiants en formation de futurs professeurs de FLE. Sur la base des résultats de travaux de recherche antérieurs sur l'acquisition du français par des apprenants avancés (Bartning/Kirchmeyer 2003, Koch/Thörle 2021), on peut supposer que ce groupe d'étudiants n'a acquis l'usage des MD que partiellement et ne dispose pas de connaissances métalinguistiques systématiques sur leur rôle en français parlé. Les objectifs de l'usage du module en cours étaient : 1. de sensibiliser les apprenants au phénomène des MD, 2. de leur faire connaître les MD en tant que classe fonctionnelle d'expressions afin qu'ils puissent les identifier dans le contexte de la conversation et de les distinguer d'autres classes de mots, 3. de leur faire reconnaître dans le corpus les fonctions décrites dans la littérature de recherche et, le cas échéant, de formuler eux-mêmes des hypothèses sur les fonctions pragmatiques dans un contexte d'interaction concret. Pour une unité d'enseignement de deux séances de séminaire, les MD *bon, enfin, quoi* ont été choisis en raison de leur fréquence relativement élevée. De plus, on peut supposer que les expressions sont familières aux apprenants, mais pas forcément connues en tant que MD. Les étudiants ont d'abord reçu une introduction à la classe

[107] Laboratoire Ligérien de Linguistique - UMR 7270 (LLL) (2023). *ESLO-FLEU : Enquêtes Sociolinguistiques à Orléans - Fle et Linguistique pour l'Enseignement Universitaire* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr v1, https://hdl.handle.net/11403/eslo-fleu/v1.

d'expression des MD (Dostie/Pusch 2007, Barme 2012) ainsi qu'une introduction aux fonctions pragmatiques et aux caractéristiques sociolinguistiques de *bon, enfin, quoi* (Beeching 2007). Sur cette base, ils ont traité la première tâche liée au corpus, qui consistait à lire et à écouter attentivement certains extraits du module ESLO « Parler jeune », à relever dans un travail en groupe toutes les occurrences de *bon, enfin, quoi* et à distinguer les usages comme MD des usages comme adjectif (*bon*), adverbe temporel (*enfin*) ou pronom interrogatif (*quoi*) (cf. les différents usages de *bon* aux lignes 3 et 6 de l'extrait 1).

1 **[AJ38] 0:01:14** donc euh du coup euh vu que bon j' avais pris quand même deux
2 **[AJ38] 0:01:17** euh enfin j' avais pris plus de cinq ans de cours de chant
3 **[AJ38] 0:01:19** je me suis dit bon c' est bon euh ça j' ai j- voilà
4 **[AJ38] 0:01:22** j' ai appris entre guillemets ce que y avait euh principalement à apprendre euh
5 **[AJ38] 0:01:27** je débla- j' ai bien développé mon appareil vocal euh
6 **[AJ38] 0:01:30** j' ai une bonne oreille enfin bon ça ça devrait rouler quoi
puis bah je continue à m' exercer

Extrait 1: ESLO2_ENT_1038_guitare, tiré du Module "Parler jeune"

Ils ont également reçu la tâche de déterminer la position des MD, de faire attention à la réalisation phonétique du marqueur *enfin*, souvent élidé (forme *'fin*) dans le module, et de noter les combinaisons des trois MD avec d'autres MD (p. ex. *enfin bon, ah bon, enfin...quoi* etc.). Lors de la deuxième séance, le groupe a découvert des approches descriptives concrètes des trois MD (basées sur Hansen 1995 pour *bon*, Bertrand/Chanet 2005 pour *enfin* et Hölker 2010 pour *quoi*) et a reçu la tâche de former sur cette base des hypothèses sur les fonctions pragmatiques de *bon, enfin* et *quoi* dans un extrait de conversation spécifique du module (« Guitare »).

L'application du module « Parler jeune » en cours de linguistique française a ensuite été évaluée à l'aide d'un questionnaire. Les réponses des dix participants à l'enquête montrent que l'authenticité des données, le support sonore et la contextualisation situative et sociale ont été évalués positivement, comme en témoigne les remarques suivantes :

« Ce qui m'a particulièrement plu était qu'on a pu travailler avec des enregistrements « authentiques ». C'était intéressant de faire des analyses par rapport aux marqueurs du discours qui peuvent avoir plusieurs sens. »

« (…) Les phénomènes linguistiques abordés ont été rendus plus compréhensibles grâce à des exemples authentiques » (Traduction)

Les difficultés rencontrées par les étudiants concernaient dans quelques cas la compréhension des extraits ainsi que la manipulation du corpus à partir du site web (nécessaire à l'époque pour accéder à l'alignement son-texte). Les activités menées en cours ainsi que les retours de étudiants ont été pris en compte pour développer une forme de présentation appropriée du module sur un site web dédié[108] ainsi que pour l'élaboration d'une proposition de cours.

[108] Site du projet : http://eslo.huma-num.fr/index.php/pagelarecherche/projets-de-l-equipe-et-sous-corpus/eslo-fleu
Accès direct au modules *Le Parler jeune* : https://lll.cnrs.fr/eslo-fleu-module-2/

Résultats : livrables

Élaboration d'un système d'annotation pour les MD dans le module « Parler jeune »

Il a été élaboré un système d'annotation qui rend compte de l'état de la question en linguistique, dont les catégories et valeurs sont néanmoins « ouvertes » du point de vue théorique et dans leur degré de spécificité accessibles pour un public universitaire d'étudiants en langue et linguistique françaises, mais pas forcément spécialisé dans la matière. Le guide d'annotation prochainement déposé sur Ortolang permet d'expliquer et d'illustrer tous les choix effectués, qu'ils soient linguistiques ou techniques, et inclut une bibliographie sélective pour chaque module.

Si le travail d'annotation n'a pas été mené en vue d'une analyse quantitative, il est néanmoins intéressant de se pencher sur le relevé du nombre des occurrences de chaque marqueur (table 1).

MD	Occurrences	MD	Occurrences
euh	448	du coup	14
mais	124	et tout	14
donc	84	genre	11
quoi	63	tout ça	9
enfin	62	attends	4
en fait	32	en même temps	3
voilà	31	alors là	2
bon	24	au final	2
après	17	bref	2
quand même	16	d'ailleurs	2
alors	14	en effet	2

table 1 : Occurrences des MD annotés dans le Module "Parler jeune"

Ces résultats doivent permettre d'orienter les choix didactiques des enseignants : ainsi, il semble pertinent d'attirer l'attention des étudiants sur des marqueurs fréquents et particulièrement polyfonctionnels comme *mais*, *donc, quoi, enfin, en fait, voilà*, conformément à l'expérimentation menée à Siegen qui s'est focalisée sur *enfin, quoi* et *bon*.

L'annotation sémantico-pragmatique permet d'orienter l'observation en pointant les MD en contexte pour suggérer une catégorisation de ces derniers et mettre en évidence leur caractère polyfonctionnel, comme dans le cas de *quoi* illustré ci-dessous (table 2).

Information	Clôture_topic		
Connexion			
Interaction	Emphase		
Formulation	Approximation	Précision	Proposition_formulation

table 2 : Domaines fonctionnels et jeu d'étiquettes pour le MD quoi

Élaboration d'un dispositif technique via un site web pour la mise à disposition du module au public ciblé

L'exploration peut se faire sous TXM (utilisateur formé) ou sur des pages web (public non expert). Pour ces deux explorations, l'accent a été mis sur l'écoute de la donnée sonore, les annotations et les métadonnées (enregistrements et locuteurs). L'interface de TXM est complexe mais permet en une requête de chercher précisément des phénomènes pour des exemples à utiliser en salle de cours. Bien qu'il soit facile de charger le corpus dans l'outil, il est plus difficile de prendre en main le système de requêtes. Pour les étudiants également, l'utilisation de logiciel est parfois à proscrire car tous n'ont pas les mêmes conditions matérielles. Ainsi il a été pensé une version HTML téléchargeable en local et également mise en ligne pour une utilisation web sans téléchargement, nécessitant une connexion internet et un navigateur. Cette version propose une visualisation des métadonnées de chaque extrait, une écoute de la donnée sonore avec sa transcription et un accès aux annotations via un bouton qui surligne les phénomènes à observer dans les différents modules sélectionnés.

Dans l'interface web, l'exploration se fait de préférence selon une approche sémasiologique : le repérage des formes et leur analyse en contexte permet d'accéder aux sens et fonctions des MD. Pour une approche onomasiologique partant des domaines fonctionnels et des étiquettes sémantico-pragmatiques pour aller vers les formes, un travail sous TXM en appui au guide d'annotation est plus indiqué.

2. Élaboration d'une proposition pédagogique

Le module est conçu comme ressource ouverte dont l'usage n'est pas prédéterminé et qui peut être exploitée selon les intérêts et besoins pédagogiques. Pour donner un exemple d'une possible exploitation didactique, une fiche pédagogique a été élaborée sur la base de l'expérimentation à l'université de Siegen (voir annexe). L'analyse de trois MD différents s'étant avérée trop vaste, les MD dans la fiche pédagogique ont été réduits à *enfin* et *quoi.*

Conclusion : perspectives

Les expérimentations didactiques ont été réalisées sur une version du sous-corpus non annotée et encore peu accessible. Il s'agira de poursuivre les expérimentations didactiques à partir de la version définitive annotée du sous-corpus, dont l'exploration en ligne est désormais facilitée suite à des aménagements techniques.

D'autres modules, en cours de publication, doivent permettre d'explorer plusieurs phénomènes comme les liaisons, les interrogatives (module *D'une situation à l'autre*), le discours rapporté direct et les clivées (module *Récit en interaction*). Des expérimentations sont en cours.

Références

Abouda, Lotfi (2022). « L'émergence du marqueur méta-discursif *du coup* : de la conséquence à l'actualisation énonciative », *Langages*, vol. 226, no. 2, 99-116.

Abouda, Lotfi & Skrovec, Marie. (2018). "Micro-diachronie d'un marqueur discursif. *En même temps* : simultanéité, coexistence, adversativité". *DIA 3 : Le français innovant*, Mar 2018, Paris, France. 245-270. ⟨halshs-03023109⟩

Bartning, Inge, Kirchmeyer, Nathalie (2003). « Le développement de la compétence textuelle à travers les stades acquisitionnels en français L2 », Acquisition et Interaction en Langue Étrangère 19, 9–39. https://doi.org/10.4000/aile.1112

Beeching, Kate (2014). « Corpora in language teaching and learning », *Recherches en didactique des langues et des cultures*, 11-1.

Bertrand, Roxane, Chanet, Catherine (2005). « Fonctions pragmatiques et prosodie de *enfin* en français spontané », *Revue de sémantique et pragmatique* 17, 41-68.

Borreguero Zuloaga, Margarita, Thörle, Britta (2016), « Introduction. Discourse Markers in Second Language Acquisition: Studies on Italian and French as L2 », *Language, Interaction and Acquisition* 7:1, 1-16.

CECRL, Conseil de l'Europe (2001), *Cadre européen commun de référence. Apprendre, enseigner, évaluer.*

Dostie, Gaétane (2004). *Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique.* De Boeck Supérieur.

Dostie, Gaetane, Pusch, Claus D. (2007), « Introduction. Les marqueurs discursifs. Sens et variation ». *Langue française* 154/2, 3–12.

Gülich, Elisabeth (1993). « Procédés de formulation et travail conversationnel : éléments d'une théorie des processus de la production discursive ». In : Hilty, Gerold (ed.) : *Actes du XXe Congrès International de Linguistique et Philologie Romanes. Université de Zurich (6-11 avril 1992).* Vol 2. Tübingen: Narr; Francke Attempto, 137-151.

Hölker, Klaus (2010). « Frz. 'quoi' als Diskursmarker », *Linguistik online* 44.

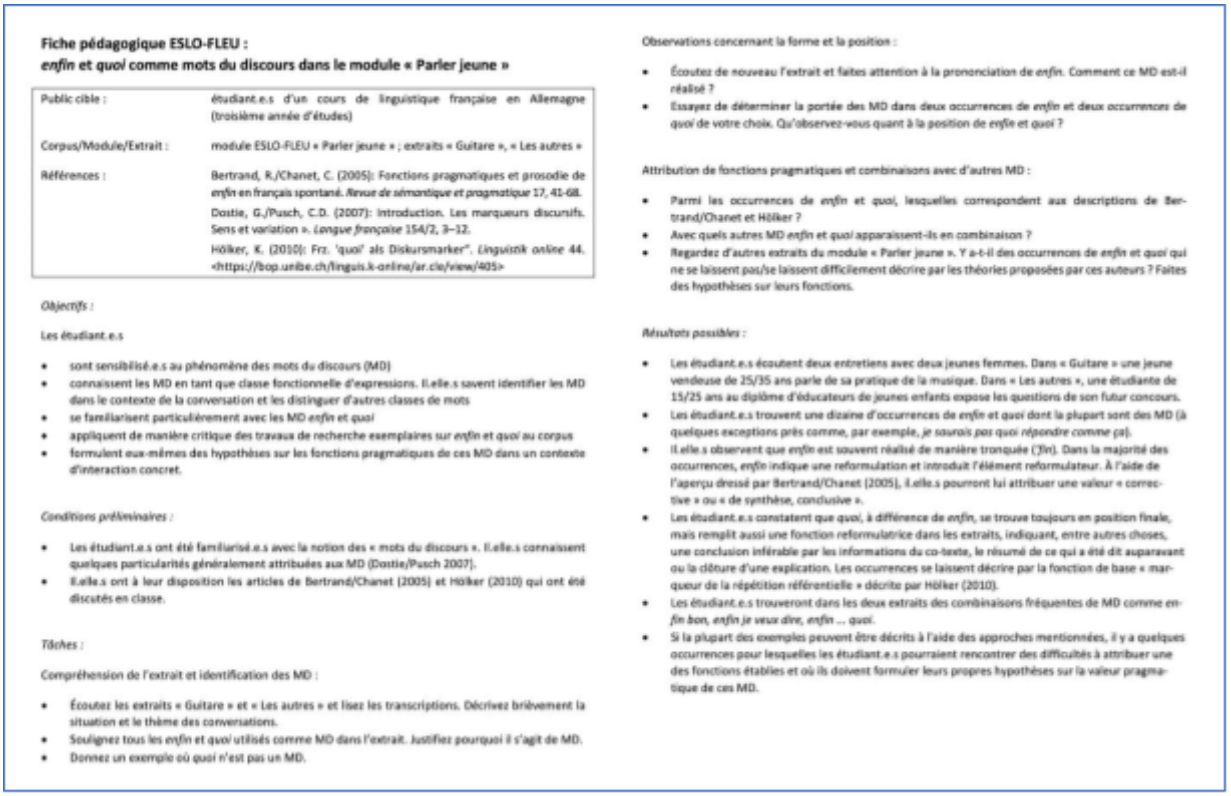
Koch, Christian, Thörle, Britta (2021). « Metadiscursive Activities in Oral Discourse Production in French L2: A Study on Learner Profiles », *Corpus Pragmatics* 5/1, 153-186.

Koch, Peter, Oesterreicher, Wulf (1990), *Gesprochene Sprache in der Romania : Französisch, Italienisch, Spanisch*, Tübingen, Niemeyer.

Skrovec, Marie, Layal Kanaan-Caillol, et Hisae Akihiro (2022). « Le marqueur *après* à l'oral : une approche micro-diachronique, variationniste et interactionnelle », *Langages*, vol. 226, no. 2, 2022, 117-131.

Annexe

Fiche pédagogique : *enfin* et *quoi* comme mots du discours dans le module « Parler jeune »



Fiche pédagogique ESLO-FLEU :
***enfin* et *quoi* comme mots du discours dans le module « Parler jeune »**

Public cible :	étudiant.e.s d'un cours de linguistique française en Allemagne (troisième année d'études)
Corpus/Module/Extrait :	module ESLO-FLEU « Parler jeune » ; extraits « Guitare », « Les autres »
Références :	Bertrand, R./Chanet, C. (2005): Fonctions pragmatiques et prosodie de *enfin* en français spontané. *Revue de sémantique et pragmatique* 17, 41-68. Dostie, G./Pusch, C.D. (2007): Introduction. Les marqueurs discursifs. Sens et variation ». *Langue française* 154/2, 3–12. Hölker, K. (2010): Frz. 'quoi' als Diskursmarker". *Linguistik online* 44. <https://bop.unibe.ch/linguis.k-online/ar.cle/view/405>

Objectifs :

Les étudiant.e.s

- sont sensibilisé.e.s au phénomène des mots du discours (MD)
- connaissent les MD en tant que classe fonctionnelle d'expressions. Il.elle.s savent identifier les MD dans le contexte de la conversation et les distinguer d'autres classes de mots
- se familiarisent particulièrement avec les MD *enfin* et *quoi*
- appliquent de manière critique des travaux de recherche exemplaires sur *enfin* et *quoi* au corpus
- formulent eux-mêmes des hypothèses sur les fonctions pragmatiques de ces MD dans un contexte d'interaction concret.

Conditions préliminaires :

- Les étudiant.e.s ont été familiarisé.e.s avec la notion des « mots du discours ». Il.elle.s connaissent quelques particularités généralement attribuées aux MD (Dostie/Pusch 2007).
- Il.elle.s ont à leur disposition les articles de Bertrand/Chanet (2005) et Hölker (2010) qui ont été discutés en classe.

Tâches :

Compréhension de l'extrait et identification des MD :

- Écoutez les extraits « Guitare » et « Les autres » et lisez les transcriptions. Décrivez brièvement la situation et le thème des conversations.
- Soulignez tous les *enfin* et *quoi* utilisés comme MD dans l'extrait. Justifiez pourquoi il s'agit de MD.
- Donnez un exemple où *quoi* n'est pas un MD.

Observations concernant la forme et la position :

- Écoutez de nouveau l'extrait et faites attention à la prononciation de *enfin*. Comment ce MD est-il réalisé ?
- Essayez de déterminer la portée des MD dans deux occurrences de *enfin* et deux occurrences de *quoi* de votre choix. Qu'observez-vous quant à la position de *enfin* et *quoi* ?

Attribution de fonctions pragmatiques et combinaisons avec d'autres MD :

- Parmi les occurrences de *enfin* et *quoi*, lesquelles correspondent aux descriptions de Bertrand/Chanet et Hölker ?
- Avec quels autres MD *enfin* et *quoi* apparaissent-ils en combinaison ?
- Regardez d'autres extraits du module « Parler jeune ». Y a-t-il des occurrences de *enfin* et *quoi* qui ne se laissent pas/se laissent difficilement décrire par les théories proposées par ces auteurs ? Faites des hypothèses sur leurs fonctions.

Résultats possibles :

- Les étudiant.e.s écoutent deux entretiens avec deux jeunes femmes. Dans « Guitare » une jeune vendeuse de 25/35 ans parle de sa pratique de la musique. Dans « Les autres », une étudiante de 15/25 ans au diplôme d'éducateurs de jeunes enfants expose les questions de son futur concours.
- Les étudiant.e.s trouvent une dizaine d'occurrences de *enfin* et *quoi* dont la plupart sont des MD (à quelques exceptions près comme, par exemple, *je saurais pas quoi répondre comme ça*).
- Il.elle.s observent que *enfin* est souvent réalisé de manière tronquée ('fin). Dans la majorité des occurrences, *enfin* indique une reformulation et introduit l'élément reformulateur. À l'aide de l'aperçu dressé par Bertrand/Chanet (2005), il.elle.s pourront lui attribuer une valeur « corrective » ou « de synthèse, conclusive ».
- Les étudiant.e.s constatent que *quoi*, à différence de *enfin*, se trouve toujours en position finale, mais remplit aussi une fonction reformulatrice dans les extraits, indiquant, entre autres choses, une conclusion inférable par les informations du co-texte, le résumé de ce qui a été dit auparavant ou la clôture d'une explication. Les occurrences se laissent décrire par la fonction de base « marqueur de la répétition référentielle » décrite par Hölker (2010).
- Les étudiant.e.s trouveront dans les deux extraits des combinaisons fréquentes de MD comme *enfin bon*, *enfin je veux dire*, *enfin … quoi*.
- Si la plupart des exemples peuvent être décrits à l'aide des approches mentionnées, il y a quelques occurrences pour lesquelles les étudiant.e.s pourraient rencontrer des difficultés à attribuer une des fonctions établies et où ils doivent formuler leurs propres hypothèses sur la valeur pragmatique de ces MD.

Monologuer dans une discussion en ligne? Profilage des interactions entre les rédacteurs de la Wikipédia

Ludovic Tanguy [1], Céline Poudat [2] et Lydia-Mai Ho-Dac [1]
[1] Laboratoire CLLE : CNRS & Université de Toulouse
[2] Laboratoire BCL : CNRS & Université Côte d'Azur
ludovic.tanguy@univ-tlse2.fr, celine.poudat@univ-cotedazur.fr, lydia-mai.ho-dac@univ-tlse2.fr

Introduction

La présente étude s'inscrit dans un travail plus général de profilage des interactions en ligne entre les rédacteurs de la Wikipédia. Ces discussions prennent place dans des espaces spécifiques associés à chacun des articles de l'encyclopédie en ligne et constituent des échanges autour des différentes décisions que nécessite l'écriture collaborative. Elles permettent notamment aux contributeurs de se coordonner pour rédiger et mettre à jour l'article, de régler leurs éventuels différends.

Sur la base d'un corpus constitué de plus de 300 000 discussions associées aux articles de Wikipédia dans sa version française, nous proposons une investigation globale des interactions et des pratiques. Cette étude se concentre sur un phénomène relativement inattendu : les monologues. À travers un examen ciblé de ce type d'interaction, nous dégageons un ensemble de pratiques qui traduisent la spécificité des discussions Wikipédia par rapport, notamment, aux forums de discussion classiques.

Vue d'ensemble des discussions Wikipédia

Notre travail se base sur un corpus de discussions élaboré à partir du dump de la Wikipédia française téléchargé en septembre 2019. Nous avons sélectionné l'ensemble des pages de discussion associées aux articles, ainsi que leurs archives. Chaque page a été segmentée en fils de discussion selon les sections délimitées par les contributeurs et chaque fil en messages sur la base des signatures, indentations et autres marques de segmentation (le format wiki utilisé permet une grande souplesse). Seules les discussions comprenant au moins un message et 2 mots ont été conservées. À chaque message est associé l'identifiant du contributeur (ou l'adresse IP des contributeurs anonymes), sa date d'écriture et bien entendu le contenu textuel. Les données sont encodées au format XML selon la norme TEI dédiée aux communications médiées par les réseaux [Beißwenger et Lüngen, 2020]. Pour la présente étude nous avons également éliminé toutes les discussions faisant intervenir un robot (en nous basant sur l'identifiant de l'auteur).

Au final, nous disposons d'un corpus de 302 475 discussions exploitables pour un total de 769 880 messages. Ces discussions sont très hétérogènes en termes de taille, durée, nombre de participants, et bien entendu de contenu. Certaines particularités de ces discussions ont déjà été décrites dans de nombreuses études s'intéressant notamment aux conflits et aux actes de dialogues [Ferschke et al., 2012, Yasseri et al., 2012, Kittur et Kraut, 2010, Viégas et al.,

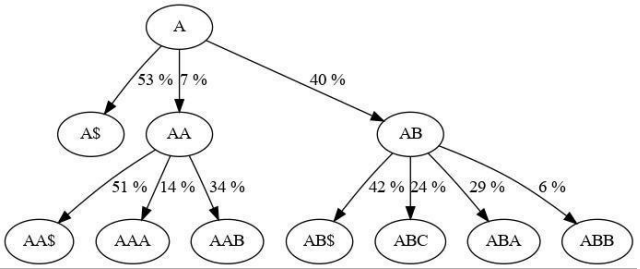
2004, Viégas et al., 2007, Stvilia et al., 2008, Wilkinson et Huberman, 2007]. Les principales caractéristiques sont résumées dans la table 1.

Caractéristique	Min.	Max.	Médiane	Moyenne
Nombre de messages par discussion	1	149	1	2,54
Nombre de participants par discussion	1	43	1	1,72
Durée (au moins 2 messages)	1 mn	16 ans	2,1 jours	184 jours

table 15. : **Caractéristiques des discussions**

Schémas d'interaction

Afin de nous concentrer sur les interactions, nous avons identifié pour chaque discussion l'ordre dans lequel les différents participants intervenaient. Indépendamment de l'identité virtuelle de chaque participant, nous désignons par A l'auteur du premier message, par B le premier contributeur (distinct de A) qui intervient, etc. La figure 1 montre comment se déploient les discussions de notre corpus en se concentrant sur les 3 premiers messages des fils (le $ indique la fin de la discussion).



: Répartition des interactions en début de fil de discussion

Comme le montre la figure 1, la majorité des fils de discussion ne contient au final pas d'échange mais un seul message posté auquel personne n'a répondu (53% des fils discussions correspondent au schéma A$, phénomène classique dans les échanges en ligne, cf. [Beaudouin et Velkovska, 1999]). Le second schéma le plus récurrent renvoie à des situations plus classiques de dialogue entre A et B (40% des interactions) - l'échange majoritaire demeurant un simple AB$.

Nous avons été particulièrement interpellés par les situations qu'on pourrait appeler monologales, qui sont de notre point de vue les moins étudiées et les moins prévisibles : les interactions démarrant par AA représentent ainsi 7% des schémas identifiés, soit 19 947 fils de discussion. Comme indiqué dans la figure 1, ce second message sera le dernier dans 51% des cas (AA$), sera suivi d'un troisième message du même auteur (14%, AAA) ou de l'intervention d'un interlocuteur (34%, AAB). Il n'est pas possible de conclure directement sur la fin d'une discussion, notamment parce que plus de la moitié des interventions n'entraînent pas de réponse. De plus, le format utilisé dans Wikipedia ne permet pas de déclarer qu'un fil est clos comme cela peut se faire sur d'autres plateformes d'échange, les discussions ayant la même durée de vie qu'une page de l'encyclopédie. Nous avons par exemple pu observer des cas où une réponse intervenait jusqu'à 15 années plus tard (soit presque l'empan temporel du corpus).

Si l'on se concentre sur les monologues purs, c'est-à-dire les fils dans lesquels un seul auteur s'est exprimé, ils représentent 57,5% des fils de discussions (174 009 sur 302 475). Parmi

ceux-ci, seuls 7% (12 023) sont de véritables monologues avec plus d'un message, et 8,5% des fils avec plusieurs messages sont des monologues (les AA$ et AAA de la figure 1). Le phénomène est donc loin d'être négligeable, sans même compter les séquences monologales au sein de discussions à plusieurs.

Nous avons décidé d'aborder l'étude des monologues par deux angles. Le premier est d'observer les situations les plus extrêmes, qui sont généralement bien plus simples à interpréter et qui permettent, au-delà de la simple anecdote, d'identifier des configurations que l'on peut par la suite retrouver avec des amplitudes plus réduites. Le second angle consiste à examiner les fils qui commencent par l'échange d'un utilisateur avec lui-même, afin d'étudier les conditions d'installation d'un monologue.

Monologues longs : étude des cas extrêmes

Nous avons donc dans un premier temps extrait les discussions monologales les plus longues : 136 fils contiennent ainsi plus de 5 messages (le plus long en contenant 28). Leur observation a permis d'identifier trois usages récurrents principaux. Le premier, que nous appellerons **tableau de bord**, correspond à l'utilisation d'un fil de discussion comme un outil de suivi des opérations réalisées ou à effectuer sur la page (vérifications, corrections, traductions etc.)[109]. Les messages de ces fils se distinguent par l'absence de marques d'interlocution. Deuxième usage, les **véritables monologues** qui se déclinent selon deux configurations puisque le monologue peut être subi ou choisi. Dans le premier cas, le monologue peut en fait être un monodialogue, dans lequel *A* tente d'entrer en contact avec un ou plusieurs autres Wikipédiens possiblement désignés explicitement qui ne lui répondent pas[110]. Ces fils contiennent d'ailleurs une proportion significative de plaintes et de reproches. Dans le second cas, *A* va écrire une série de messages dont l'enchaînement logique correspond à une réflexion ou une investigation[111]. Bien loin de la liste, ces messages sont de véritables textes exposant des faits, des interprétations, souvent argumentés et traduisant l'évolution du point de vue de *A*. On se rapproche alors de certaines formes de journaux extimes et on y note l'absence de marque d'une volonté interlocutrice : pas d'appel à un tiers ou à une communauté anonyme pour obtenir un avis, un complément ni même un assentiment. Le dernier usage observé correspond à des cas de **séries pures**, qui correspondent souvent à des listes d'items, rajoutés progressivement. Un cas extrême de série pure est celui d'un utilisateur qui, à de nombreuses reprises dans les pages de discussion associées à un acteur de cinéma non francophone, établit la liste des acteurs français qui ont assuré son doublage, avec un message pour chaque film - sans pour autant jamais éditer la page Article correspondante[112].

[109] Voir par exemple : https://fr.wikipedia.org/wiki/Discussion:Panth%C3%A9on_pyr%C3%A9n%C3%A9en#Faux_dieux_pyr%C3%A9n%C3%A9ens,_fausses_localisations,_etc.

[110] Voir par exemple : https://fr.wikipedia.org/wiki/Discussion:Virginie_Grimaldi#%C2%AB_Chick_lit_%C2%BB_et_%C2%AB_feel_good_%C2%BB

[111] Voir par exemple : https://fr.wikipedia.org/wiki/Discussion:L%27H%C3%B4tel#%C2%AB_L%E2%80%99H%C3%B4tel_%C2%BB_situ%C3%A9_%C3%A0_l%E2%80%99emplacement_de_la_r%C3%A9sidence_de_Marguerite_de_France_?

[112]4 Voir par exemple : https://fr.wikipedia.org/wiki/Discussion:Ben_Johnson_(acteur)#Voix_Fran%C3%A7aise

Monologues en début de discussion : pourquoi se répondre à soi-même ?

Afin de comprendre comment un monologue s'installe et se déploie, nous avons examiné un échantillon de 100 commencements de situations monologales, c'est-à-dire 100 fils dans lesquels l'auteur du premier message est le premier à intervenir à nouveau dans la discussion (schémas commençant par AA dans la figure 1), sélectionnés aléatoirement. Nous avons tenté de spécifier la fonction de chaque second message indépendamment de la suite du fil. Au final, plus de la moitié des débuts de monologues renvoient à deux cas de figure dominants : dans le premier cas, on observe un schéma **suggestion-action** (28%) dans lequel *A* suggère ou demande explicitement une action sur l'article qu'il indique ensuite avoir réalisée[113]. Le fil de discussion peut s'arrêter là ou se poursuivre, notamment par l'intervention d'un interlocuteur en désaccord avec l'action effectuée. Soulignons ici que la notification d'une *action* est propre au travail collaboratif visé par cet espace de discussion. Suggérer, demander, annoncer ou valider une modification de la page sont des motivations très courantes pour les discussions (*cf* la catégorie *explicit performative* de [Ferschke *et al.*, 2012] qui correspond à environ 60% des messages qu'ils ont annotés). Cette catégorie semble aller de pair avec celles des **tableaux de bord** ou des **séries pures** (11% à elles deux), que nous avions identifiées dans les monologues longs : dans ce cas, les deux messages du même auteur forment une liste et sont généralement suivis d'autres items du même type. Ces listes peuvent concerner des remarques sur l'état de l'article, des problèmes identifiés, des actions à effectuer etc. On est ici dans une autre conception de la page de discussion Wikipédia qui s'apparenterait davantage à un tableau de suivi qu'à un outil de dialogue, ce qui nous semble en partie lié au format Wiki. En effet, les pages de discussion Wikipédia s'éditent comme l'article et prennent donc plus facilement la forme d'un texte écrit, qui peut avoir une certaine cohésion par rapports aux plateformes de discussion de type forum qui ne permettraient pas l'émergence de telles formes.

Une autre situation également propice à l'établissement d'un monologue a trait au second cas dominant relevé : la **complétion** (27%), dans lequel *A* complète sa remarque ou sa question initiale avec de nouvelles informations ou précisions[114]. Cet ajout peut s'accompagner d'une relance (ciblée ou non) afin d'obtenir une réponse ou une réaction tandis que dans d'autres situations les deux messages forment un raisonnement qui évoque une sorte de pensée à voix haute (avancée d'arguments contradictoires, évolutions de la position par rapport sur une constatation précédente, etc.). La discussion est alors rarement close et peut se poursuivre, prenant également potentiellement la forme d'un texte fragmentaire. Nous retrouvons ici les **véritables monologues** mentionnés supra.

Les autres situations relevées sont plutôt du côté de l'interaction et sont donc moins propices à l'installation d'un monologue. Outre la **relance** pure[115] (ciblée ou non) qui serait plus marginale (2%), nous avons observé deux cas qui semblent être des paires d'actes de langage associés, le deuxième message clôturant généralement le fil : (i) **message-rectification** (12%) : *A* rectifie son message initial, généralement en admettant une erreur ou en ajoutant

[113] Voir par exemple : https://fr.wikipedia.org/wiki/Discussion:Romani#Nombre_de_locuteurs_par_pays

[114] Voir par exemple : https://fr.wikipedia.org/wiki/Discussion:S%C3%A9isme_et_tsunami_de_2004_dans_l%27oc%C3%A9an_Indien#Reprise_article_de_qualit%C3%A9?

[115] Voir par exemple : https://fr.wikipedia.org/wiki/Discussion:Baruch_Goldstein#Comparatif_de_la_version_en_cours_et_la_version_propos%C3%A9e_par_Parmatus

une information qui le rend caduc, ce qui met généralement fin à la discussion[116] ; et (ii) **question-réponse** (2%) : *A* répond lui-même à la question qu'il avait posée à la communauté, ce qui là aussi termine généralement l'échange[117].

Enfin, une particularité des discussions Wikipédia par rapport à d'autres types d'interactions en ligne tient à leur adossement aux articles en cours de rédaction qui correspondent en quelque sorte à un monde extralinguistique de référence ; ainsi les actions sur l'article peuvent être annoncées et commentées par leur auteur comme nous l'avons vu. Mais un autre phénomène nous a intéressé, à savoir la **réaction** de *A* à un événement (16%), par exemple une édition de l'article par un tiers. Cette réaction peut prendre différentes formes : celle du remerciement ou de l'encouragement[118] (rare), de la critique[119] (plus fréquent) ou encore de la demande d'information ou de complément[120]. À noter que ces situations sont souvent complexes à suivre, car le contenu du deuxième message n'est pas toujours explicite et peut faire référence à d'autres espaces de discussion (un autre article ou la page personnelle d'un contributeur). Ces configurations sont le produit de la complexité d'un écosystème communicationnel multicanal comme celui de la Wikipédia, mais font aussi écho à des modalités d'interaction dans des contextes d'interaction autour de manipulations partagées entre deux interlocuteurs (par exemple [Mondada, 2006, Koester, 2006]).

Ce dernier phénomène reste à explorer de façon plus systématique. Plus globalement nombre des situations spécifiques aux discussions autour des pages Wikipédia que nous avons mises au jour dans l'étude des monologues peuvent se retrouver au sein de situations plus variées entre différents locuteurs.

Bibliographie

Beaudouin et Velkovska (1999). Beaudouin, V. et Velkovska, J. (1999). Constitution d'un espace de communication sur internet (forums, pages personnelles, courrier électronique...). *Réseaux. Communication-Technologie-Société*, 17(97):121–177.

Beißwenger et Lüngen (2020). Beißwenger, M. et Lüngen, H. (2020). CMC-core : a schema for the representation of CMC corpora in TEI. *Corpus*, 20.

Ferschke *et al.* (2012). Ferschke, O., Gurevych, I. et Chebotar, Y. (2012). Behind the article : Recognizing dialog acts in Wikipedia talk pages. *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786. Association for Computational Linguistics.

116 Voir par exemple : https://fr.wikipedia.org/wiki/Discussion:Marie-Th%C3%A9r%C3%A8se_de_France_(1778-1851)#Ses_m%C3%A9moires

117 Voir par exemple : https://fr.wikipedia.org/wiki/Discussion:Comt%C3%A9_de_Hudson#Union_City,_une_ville_du_comt%C3%A9,_est_la_plus_peupl%C3%A9e_des_%C3%89tats-Unis.

118 Voir par exemple : https://fr.wikipedia.org/wiki/Discussion:Art_tib%C3%A9tain

119 Voir par exemple : https://fr.wikipedia.org/wiki/Discussion:Musique_classique/archive1#Ah,_a_propos_de_musique_%22savante%22

120 Voir par exemple https://fr.wikipedia.org/wiki/Discussion:Harry_Potter_et_le_Prince_de_sang-m%C3%AAl%C3%A9_(film)#Incoh%C3%A9rences_majeures_par_rapport_au_livre

Kittur et Kraut (2010). Kittur, A. et Kraut, R. E. (2010). Beyond Wikipedia : coordination and conflict in online production groups. *In Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 215–224. ACM.

Koester (2006). Koester, A. (2006). *Investigating workplace discourse*. Routledge.

Mondada (2006). Mondada, L. (2006). Interactions en situations professionnelles et institutionnelles : de l'analyse détaillée aux retombées pratiques. *Revue française de linguistique appliquée*, 11(2):5–16.

Stvilia *et al.* (2008). Stvilia, B., Twidale, M. B., Smith, L. C. et Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001.

Viégas *et al.*, (2004). Viégas, F. B., Wattenberg, M. et Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582. ACM.

Viégas *et al.* (2007). Viégas, F., Wattenberg, M., Kriss, J. et van Ham, F. (2007). Talk Before You Type : Coordination in Wikipedia. *In 40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007*, pages 78–78.

Wilkinson et Huberman (2007). Wilkinson, D. M. et Huberman, B. A. (2007). Cooperation and Quality in Wikipedia. *In Proceedings of the 2007 International Symposium on Wikis*, WikiSym '07, pages 157–164, New York, NY, USA. ACM.

Yasseri *et al.* (2012). Yasseri, T., Sumi, R., Rung, A., Kornai, A. et Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PloS one*, 7(6):e38869.

Diffusion des innovations lexicales sur Twitter : description et prédiction de l'influence de la position des locuteurs dans le réseau

Louise Tarrade [1], Jean-Pierre Chevrot [2], Jean-Philippe Magué [1]
[1] Laboratoire ICAR (UMR 5191), École Normale Supérieure de Lyon
[2] Laboratoire LIDILEM (EA 609), Université Grenoble Alpes
louise.tarrade@ens-lyon.fr, jean-pierre.chevrot@univ-grenoble-alpes.fr, jean-philippe.magueé@ens-lyon.fr

Introduction[121]

L'étude de la variation et du changement linguistique est au cœur de la sociolinguistique variationniste. Des théories majeures sur ce sujet ont émergé de ce domaine : parmi celles-ci, les propositions de Milroy & Milroy (1985) sur l'importance des liens faibles dans l'introduction des innovations, inspirées par le sociologue des réseaux Granovetter (1973) qui a mis en évidence la fonctionnalité de ce type de connexions. Ainsi Milroy & Milroy (1985) ont défini les innovateurs, les personnes apportant l'innovation dans leur communauté, comme ayant des positions périphériques à leur communauté avec de nombreux liens faibles à l'intérieur et à l'extérieur de celle-ci. De même, ils ont mis en évidence que pour qu'une variante s'établisse au sein d'une communauté, une condition nécessaire est qu'elle ait été préalablement adoptée par des personnes à la fois centrales et bien ancrées dans celle-ci. De son côté, Labov (2001) décrit les leaders du changement linguistique comme des personnes à la fois très centrales à leur communauté mais disposant également de nombreux liens à l'extérieur de celle-ci. Cependant, ces études portaient sur des variantes phonétiques, étaient menées à l'échelle d'une centaine d'individus, et souvent en synchronie.

La sociolinguistique computationnelle (Nguyen et al., 2016) permet d'échapper aux limites imposées par l'enquête de terrain, et d'étudier diachroniquement une variété de langue à mi-chemin entre l'écrit et l'oral, très propice à la variation et à l'innovation. À l'instar de la sociolinguistique variationniste traditionnelle, elle s'est donc en partie attelée à étudier le changement linguistique et son processus de diffusion. Par exemple, des simulations multi-agents ont permis de mettre en évidence l'importance des membres périphériques dans l'apport d'innovations et celle d'individus au centre de sous-groupes denses dans l'apparition de normes (Fagyal et al., 2010). Au niveau des médias sociaux, une attention particulière a été portée à l'importance des liens faibles dans l'apport des innovations et à l'influence des liens forts, sur des corpus issus de Facebook (Bakshy et al., 2012), Twitter (Goel et al., 2016), ou Reddit (Del Tredici & Fernández, 2018), ou encore à l'influence de la structure du réseau sur la diffusion des innovations (Zhu & Jurgens, 2021 ; Würschinger, 2021). Cependant, ce changement d'échelle questionne les notions traditionnelles de réseau, de lien et de communauté. De plus, le schéma global de l'influence de la structure du réseau aux

[121] Les auteurs remercient le LABEX ASLAN (ANR-10-LABX-0081) de l'Université de Lyon pour son soutien financier dans le cadre du programme français "Investissements d'Avenir" géré par l'Agence Nationale de la Recherche (ANR).

différentes étapes de diffusion des innovations ainsi que les paramètres qui influent sur le succès ou l'échec de ces dernières restent encore flous.

En nous basant sur la trajectoire en forme de S des innovations réussies (Blythe & Croft, 2012 ; Fagyal et al., 2010 ; Rogers, 2003), nous chercherons à caractériser les sous-populations d'utilisateurs qui s'emparent des innovations aux différentes étapes de leur diffusion. Plus précisément, il s'agira d'essayer de dégager un schéma général de la façon dont le profil des utilisateurs évolue au fur et à mesure de la diffusion d'une innovation lexicale. En comparaison avec des innovations dont la trajectoire aboutit à l'échec de leur établissement dans la communauté linguistique, nous nous demanderons qui sont les acteurs du changement et si leur position dans le réseau aux différentes phases entrave ou participe à la diffusion de ces innovations.

Corpus et méthodologie

Corpus

Nous nous appuyons pour ce travail sur un corpus d'environ 650 millions de tweets en français rédigés de 2007 à début 2019, et provenant d'environ 2,5 millions d'utilisateurs. Une collecte initiale de 170 millions de tweets produits entre 2014 et 2017 dans les fuseaux horaires GMT et GMT+1 constitue le socle de ce corpus (Abitbol et al., 2018). Celui-ci a été complété dans un second temps par une récupération des derniers tweets des utilisateurs à l'aide de l'API Twitter, puis filtré en fonction de la langue et du client utilisé afin de ne garder que les tweets en français et d'éliminer autant que possible les tweets provenant de bots.

En parallèle a été récupéré pour chaque utilisateur l'ensemble de ses followees, c'est-à-dire des personnes qu'il suit. À partir de ces informations, nous avons reconstitué le réseau des utilisateurs de notre corpus, dont les nœuds sont les utilisateurs, qui peuvent avoir des liens entrants (followers) et sortants (followees) internes, c'est-à-dire que nous considérons seulement les liens entre les utilisateurs de notre corpus. Nous obtenons finalement un réseau dirigé et statique de plus de 2,5 millions de nœuds et 300 millions de liens.

Méthodologie

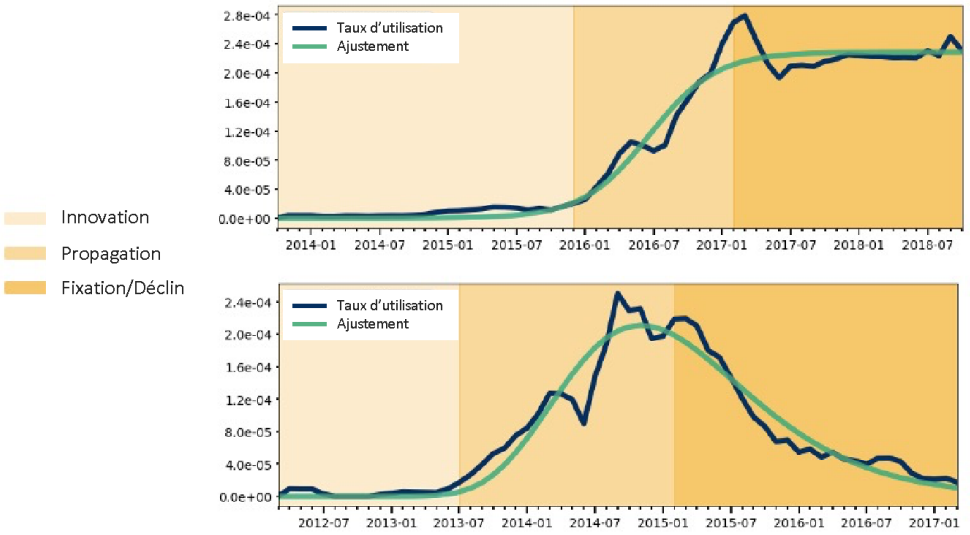
À partir du réseau modélisé des utilisateurs du corpus, nous avons distingué les communautés en nous appuyant sur le principe de l'algorithme de Louvain (Blondel et al., 2008), puis nous avons caractérisé chaque utilisateur[122] en fonction des mesures suivantes :

- le coefficient de clustering local, qui rend compte pour un utilisateur du degré d'ouverture de son réseau ;
- le score de PageRank, qui est un indicateur du prestige d'un individu et va dépendre à la fois du nombre de ses liens entrants mais également de si ces liens entrants ont eux-mêmes un score de PageRank élevé ;
- la centralité d'intermédiarité, qui mesure à quel point l'utilisateur est central à sa communauté et fait office de "pont" au sein de celle-ci ;

[122] Au regard de la taille conséquente du réseau, la modélisation du graphe ainsi que les calculs des différentes variables de réseau - à l'exception de la proximité avec l'extérieur de la communauté - ont été effectués à l'aide de la librairie Networkit (Staudt et al., 2014).

- la proximité avec les autres communautés, qui correspond au nombre de « pas » moyen nécessaire pour un utilisateur avant de pouvoir atteindre une autre communauté.

Dans un précédent travail (Tarrade et al., 2022) nous avons détecté les innovations lexicales apparues dans le corpus de tweets entre mars 2012 et février 2014. Plus précisément, nous avons sélectionné tous les mots[123] apparus pour la première fois au cours de cette période et avons conservé les mots dont la trajectoire d'utilisation sur 5 ans s'ajustait à une courbe logistique ou à une courbe gaussienne. Nous les avons ainsi catégorisés en tant que changement ou buzz selon si leur taux d'utilisation réussissait à se stabiliser au fil du temps ou non (figure . 1). Nous avons ensuite délimité de façon automatique leurs trois phases de diffusion, à savoir innovation, propagation, puis fixation pour les changements ou déclin pour les buzz. Nous avons ainsi réussi à identifier 141 changements et 251 buzz.



Deux exemples d'ajustement : « malaisante » (changement, en haut) et « sweg » (buzz, en bas)

Pour le travail présenté ici, nous avons ajouté une condition contrôle composée de 200 mots utilisés entre février 2013 et janvier 2018, dont le taux d'utilisation est stable tout au long de cette période, et dont la distribution en termes de nombre d'utilisateurs est similaire à celle des innovations lexicales que nous avons préalablement identifiées.

Afin de dégager un schéma global de diffusion de ces innovations lexicales, nous avons caractérisé chaque innovation, à chacune de ses phases de diffusion, en fonction des caractéristiques de réseau des utilisateurs l'ayant employée pour la première fois durant la phase observée. Chaque innovation et chaque mot contrôle se voit ainsi attribuer une valeur pour les quatre variables de réseau décrites précédemment. Nous comparons ensuite les distributions de ces valeurs pour les différentes catégories de mot, afin de déterminer si et comment elles diffèrent à chacune de ces phases.

Pour confirmer nos observations, nous tentons ensuite de prédire le destin des innovations lexicales avant que leur trajectoire ne se stabilise ou ne décline, soit dès la phase d'innovation ou de propagation. Pour se faire, nous entraînons un modèle de régression logistique[124] sur les

[123] C'est à dire toute suite de caractères alphanumériques pouvant contenir une apostrophe ou un tiret.
[124] Nous utilisons pour cela l'algorithme de classification par régression logistique de la librairie Scikit-learn.

innovations lexicales de notre jeu de données pour faire de la classification binaire : la variable à prédire est le type d'innovation lexicale (changement ou buzz), et les variables explicatives sont les valeurs médianes de l'ensemble des utilisateurs de chaque mot pour chacune des variables de réseau.

Résultats

Dans un premier temps, nous avons pu établir que les innovations lexicales sont utilisées pour la première fois par des utilisateurs aux caractéristiques de réseau relativement similaires, que ces innovations connaissent plus tard le succès ou l'échec. Contrairement à ce que nous pouvions attendre, elles ne sont pas utilisées lors de la phase d'innovation par des individus dont le réseau personnel est plus ouvert ou plus fermé que la moyenne. Ceux-ci ont la possibilité d'être plus vite en contact avec d'autres communautés, sans être ni centraux dans leur propre communauté, ni prestigieux au sein du réseau global. Un profil des premiers adoptants des innovations lexicales (les innovateurs) se dessine donc dès la première phase de diffusion, mais sans que celui-ci ne se distingue réellement d'un type d'innovation à l'autre - buzz ou changement.

Mais le destin des innovations lexicales se joue en phase de propagation. Lors de cette phase, les changements sont caractérisés par des adoptants plus prestigieux que ceux des buzz, à la fois centraux à leur communauté, mais également situés plus à proximité des autres communautés. À l'inverse, les buzz ne réussissent pas à atteindre des utilisateurs centraux ou ayant une proximité avec des utilisateurs extérieurs à leur communauté, ce qui entrave à priori leur diffusion dans le réseau global.

Cette configuration de paramètres distinguant les buzz des changements est confirmée par les résultats de la prédiction par régression logistique obtenus en phase de propagation, avec un taux de précision de plus de 80%. Ces résultats esquissent un schéma général de diffusion du changement linguistique en fonction des positions des locuteurs dans le réseau.

En conclusion, nous discuterons de leur adéquation avec les conclusions des précédentes études menées en sociolinguistique variationniste, notamment en ce qui concerne le profil des innovateurs décrit par Milroy & Milroy (1985) et celui des meneurs du changement par Labov (2001).

Références bibliographiques

Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P., & Fleury, E. (2018). Socioeconomic Dependencies of Linguistic Patterns in Twitter : A Multivariate Analysis. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 1125-1134. https://doi.org/10.1145/3178876.3186011

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. *Proceedings of the 21st International Conference on World Wide Web*, 519–528.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Blythe, R. A., & Croft, W. (2012). S-CURVES AND THE MECHANISMS OF PROPAGATION IN LANGUAGE CHANGE. *Language*, *88*(2), 269–304. JSTOR.

Del Tredici, M., & Fernández, R. (2018). The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities. *ArXiv:1806.05838 [Cs]*. http://arxiv.org/abs/1806.05838

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L., & Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua*, *120*(8), 2061–2079. https://doi.org/10.1016/j.lingua.2010.02.001

Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F., & Eisenstein, J. (2016). The social dynamics of language change in online networks. *International Conference on Social Informatics*, 41–57.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, *21*(2), 339–384.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational Sociolinguistics: A Survey. *ArXiv:1508.07544 [Cs]*. http://arxiv.org/abs/1508.07544

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed). New York: Free Press.

Staudt, C. L., Sazonovs, A., & Meyerhenke, H. (2014*). NetworKit: A Tool Suite for Large-scale Complex Network Analysis*. https://doi.org/10.48550/ARXIV.1403.3005

Tarrade, L., Magué, J.-P., & Chevrot, J.-P. (2022). Detecting and categorising lexical innovations in a corpus of tweets. *Psychology of Language and Communication*, *26*(1), 313-329. https://doi.org/10.2478/plc-2022-15

Würschinger, Q. (2021). Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter. *Frontiers in Artificial Intelligence*, *4*, 648583. https://doi.org/10.3389/frai.2021.648583

Zhu, J., & Jurgens, D. (2021). The structure of online social networks modulates the rate of lexical change. *ArXiv Preprint ArXiv:2104.05010*.

L'annotation de corpus : une démarche pertinente pour évaluer la qualité textuelle ? L'exemple de l'outil Inception

Sonia Tesson
[1] Laboratoire LIDILEM, Université GRENOBLE-ALPES
sonia.tesson@univ-grenoble-alpes.fr

Introduction

Aujourd'hui encore, et peut-être plus que jamais, la capacité à rédiger des textes clairs et efficaces constitue une compétence fort recherchée, et ce à différents niveaux. De nombreuses études montrent en effet les différents moments de la vie lors desquels on est évalué, voire sélectionné, selon notre degré de maitrise de cette compétence rédactionnelle (Bautier & Rayou, 2009; Delarue-Breton & Bautier, 2019; Morinet, 2012), cela depuis des décennies, et au-delà des frontières de la France (Odell, 1981 ; Sheils, 1975). Le besoin est tel que l'enseignement de l'écrit prend une place prépondérante à l'université, non seulement en licence (Arrêté du 30 juillet 2018 relatif au diplôme national de licence, s. d.), mais aussi dans les niveaux supérieurs, quelle que soit la carrière visée par l'étudiant (Rinck, 2022; Rot et al., 2014). Le défi à relever par l'université est toutefois de taille car l'accompagnement nécessaire à chacun, forcément « sur-mesure » pour qu'il soit adapté aux productions textuelles uniques qui sont produites par les étudiants, semble aux antipodes de ce que le contexte permet réellement, avec un taux d'encadrement nettement plus faible que dans le secondaire.

Plusieurs outils tels que le projet Voltaire ou encore écri+ (https://ecriplus.fr/, ANR-17-NCUN-0015) sont, en France, nés notamment de cette tension et de ce besoin d'évaluer massivement les compétences rédactionnelles des étudiants : l'objectif est de poser un diagnostic le plus précis possible de façon à adapter l'enseignement qui viendra en aval. Parmi les forces incontestables de la plateforme écri+ (https://app.tests.ecriplus.fr/connexion), dont l'objectif est notamment de permettre l'évaluation et la certification des étudiants sur leur français écrit à partir de questions fermées, on signalera d'une part la facilité d'appropriation qu'elle permet aux étudiants qui peuvent s'en emparer en toute autonomie, mais aussi la récupération massive de données d'apprentissage que la plateforme donne à ses utilisateurs enseignants, ou encore le caractère authentique des textes à partir desquels les questions sont construites. Tous ces éléments n'empêchent toutefois pas d'insister « sur le fait que s'entrainer à évaluer des énoncés et à exercer sa vigilance ne garantit pas un transfert en situation de production » (Rinck, 2022, p. 111), et les recherches sur les processus cognitifs impliqués dans la rédaction ne manquent pas d'étayer cette assertion (voir notamment (Olive & Piolat, 2003).

Dans ce contexte, l'enjeu sous-jacent se résume donc au constat que Charney faisait il y a déjà presque quarante ans, à savoir de trouver un moyen de déjouer la difficulté à construire une méthode d'évaluation qui soit à la fois fiable (*reliable* dans le texte d'origine) et valide (*valid*

dans le texte), qui permette donc, en d'autres termes, d'évaluer précisément ce que l'on souhaite évaluer, et de le faire sans qu'il y ait de grandes variations dans la notation.

Dans le cadre de ces Journées de la Linguistique de Corpus, nous souhaitions étudier une démarche où l'évaluation des textes ne serait pas automatisée et donc confiée à un outil informatique, mais simplement assistée d'un outil d'annotation de corpus : le but est de voir si cette démarche permet d'avoir un bénéfice en termes de fiabilité d'évaluation. Dans le cas présent, nous avons choisi de tester le logiciel Inception, d'une part parce qu'il permet une annotation collaborative et le calcul automatique de l'accord inter-annotateurs travaillant sur un même projet, et d'autre part pour la facilité de prise en main qu'il permet à ses utilisateurs. En nous appuyant sur le jeu d'étiquettes que nous sélectionnons ou que nous créons spécifiquement pour notre tâche dans notre environnement de travail, nous souhaitons augmenter le score de convergence inter-annotateurs et obtenir ainsi une catégorisation moins flottante des « zone[s] de révision » ou encore des « zones de fragilité » (Rinck, 2022) à faire travailler aux étudiants.

L'enjeu de cette communication est d'une part de mettre à disposition des praticiens en charge de cours d'accompagnement à l'écrit universitaire une méthode d'évaluation d'écrits d'étudiants avec le logiciel Inception facilement réutilisable ; cette démarche nous semble d'autre part permettre de nous insérer dans le projet de changement d'échelle des entreprises de constitution de corpus utiles à la didactique que M.-L. Elalouf appelait de ses vœux (Elalouf, 2011), en participant à la « cartographie des compétences scripturales en construction » (Doquet & Ponton, 2021) à l'université.

La question à laquelle il s'agit de répondre est donc la suivante : Inception constitue-t-il une aide permettant une évaluation manuelle fiable des productions textuelles de nos étudiants ?

Corpus et méthodologie

Objectifs

Notre démarche consiste, par conséquent, en une analyse comparative testant la fiabilité d'une méthode d'évaluation de textes effectuée à l'aide de l'outil d'annotation de corpus Inception. Rappelons d'abord que, selon Charney, toute mesure doit nécessairement répondre à deux critères : tout d'abord, pour être « valide », un test devra mesurer ce qu'il est censé tester, et non des connaissances ou habiletés connexes ; par ailleurs, on considérera notre test « fiable » dès lors que le score qui sera donné à telle compétence évaluée sera stable : si l'on devait reproduire la notation, l'étudiant devrait obtenir sensiblement la même note, toutes choses étant égales par ailleurs (Charney, 1984) :

> *As in other areas of research and testing, these are the two necessary criteria for any measurement. A reliable measurement is capable of replication under equivalent conditions. So, a reliable method of assessing writing ability would yield a consistent judgment of a student's abilities if applied again, all else being equal. A valid measurement assesses what it claims to assess. So, a valid writing assessment would be sensitive to a writer's "true" abilities. Time and again, the methods that have been employed to measure writing ability have been criticized as either unreliable or invalid.*

Deux variables principales peuvent faire varier le niveau de fiabilité d'un test (pour un recensement plus complet, voir l'intéressante étude de (Suchaut, 2008)). Le temps tout

d'abord : pour qu'un test soit fiable, on cherche à trouver une méthode d'évaluation qui ne produira pas des scores différents si une même personne venait à évaluer la même copie à deux instants différents (l'intervalle pouvant être de plusieurs mois, voire de plusieurs années). Ensuite, on peut observer des différences de scores du fait des évaluateurs : on recherchera donc une méthode permettant de garantir que les scores attribués seront sensiblement les mêmes, quelle que soit la personne qui en soit à l'origine.

Dans notre démarche, c'est cette seconde variable que nous étudions. Par conséquent, les objectifs de cette communication sont les suivants :

1. Etablir un recensement des erreurs perçues par les annotateurs avec et sans Inception, pour ensuite déterminer quels sont le nombre et le type d'erreurs communes aux deux méthodes d'évaluation, puis repérer, le cas échéant, les erreurs qui seraient spécifiques à l'une ou l'autre méthode ;
2. Calculer et comparer l'accord inter-annotateurs (désormais AIA) dans les deux modalités d'évaluation ;
3. Comparer le temps d'évaluation moyen propre à chaque méthode d'évaluation.

Corpus

Pour cette étude, nous avons utilisé un sous-corpus constitué (Vaguer, 2007) composé de trente-trois textes d'étudiants inscrits en première année de licence d'histoire pour la moitié d'entre eux, et de de sciences de l'éducation pour l'autre moitié. Les deux filières sont également représentées dans chacun des deux corpus. De façon majoritaire, tous les étudiants volontaires de cette étude sont nés en 2003 ou en 2004, à l'exception d'une personne qui est née en 1991 et dont le texte a été intégré au corpus 1. Les données concernant le sexe des participants n'ont pas été recueillies, mais comme on sait que les participants étudiant les sciences de l'éducation étaient très majoritairement des femmes lors du recueil, et qu'il y avait également des femmes dans la filière d'histoire participant à notre étude, on peut en conclure que notre panel est majoritairement féminin.

Ce corpus d'étude est divisé en deux de façon à ce que l'étude des deux méthodes d'évaluation soit possible :

- Corpus 1, composé des textes T1 à T16 ;
- Corpus 2, composé des textes T17 à T33.

Les évaluateurs participant à notre recherche sont au nombre de 6, et repartis dans deux groupes distincts :

- Groupe 1 : 3 évaluateurs sont chargés d'évaluer le corpus 1 sans Inception, puis le corpus 2 avec Inception ;
- Groupe 2 : 3 évaluateurs sont chargés d'évaluer le corpus 2 sans Inception, puis le corpus 1 avec Inception.

Ces 6 évaluateurs ont tous une formation ou bien en sciences du langage, ou bien en lettres modernes, ainsi qu'une expérience d'enseignement ; on propose toutefois un accompagnement individuel si les catégories d'erreurs à repérer n'étaient pas claires. On informe par ailleurs tous les participants en début de protocole qu'ils auront à évaluer deux corpus, l'un sans outil, l'autre avec Inception, et tous savent que le but de l'étude est d'étudier

le degré de fiabilité de l'outil pour l'évaluation de textes d'étudiants (sans connaitre, toutefois, les critères précis sur lesquels l'étude se concentre).

Lors de la phase de recueil, tous les textes du corpus ont été rédigés pour répondre à la consigne suivante (consigne issue de l'étude de (Boch et al., 2016)) :

> *Lors d'une enquête réalisée auprès de consommateurs, Alain expose son point de vue, que vous allez entendre dans un document sonore.*
>
> *Dans un texte rédigé, vous indiquerez en premier lieu quelle position défend Alain, puis quels arguments il utilise pour ce faire. Votre texte sera d'une longueur d'environ 15 lignes (soit entre 205 et 225 mots). NotePad++ peut vous aider à compter le nombre de mots (dans les statistiques de votre fichier).*
>
> *Remarque : les mots élidés (d', l', n', etc.) comptent pour un mot.*
>
> *** Attention : il ne s'agit pas de donner votre avis sur la question.*
>
> *Votre texte sera évalué sur le respect des consignes suivantes :*
>
> *- texte rédigé ;*
> *- nombre de mots compris dans la fourchette indiquée ;*
> *- indication de la question faisant débat et du positionnement d'Alain à ce sujet ;*
> *- présentation logique des arguments d'Alain pour construire son positionnement ;*
> *- qualité de la langue (orthographe, syntaxe, lexique, ponctuation).*

Méthodologie

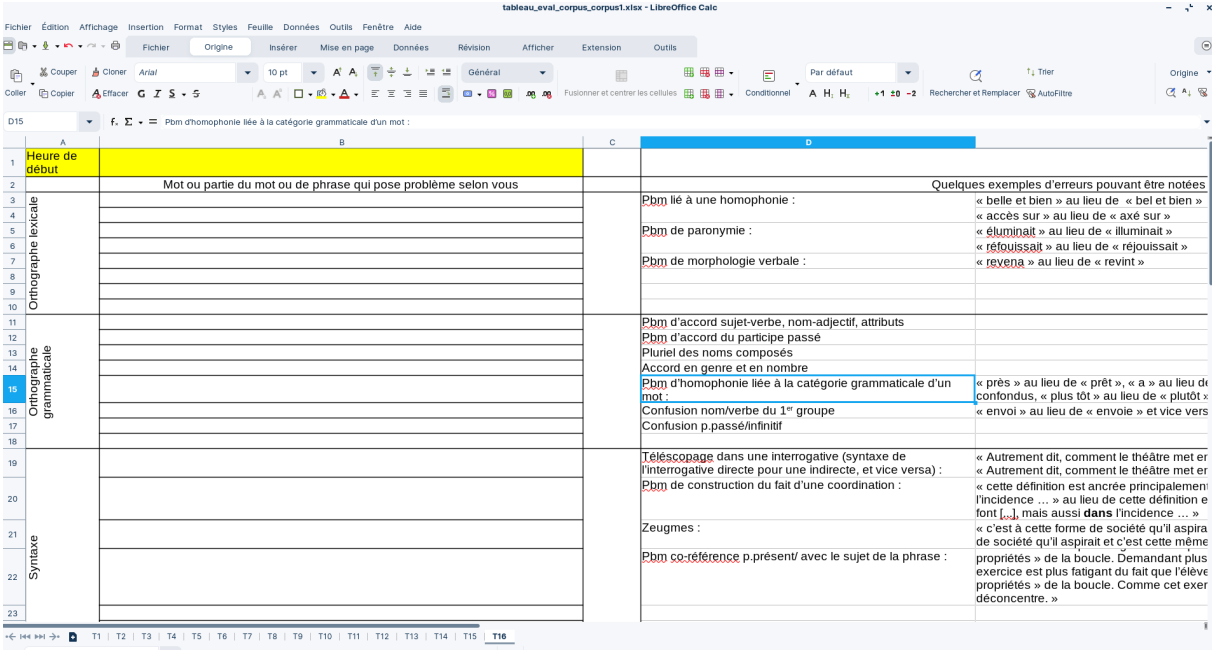
Protocole d'évaluation

Dans le cadre de cette recherche, nous avons souhaité concentrer notre attention sur les points de langue posant le plus couramment des problèmes aux étudiants. Nous avons ainsi élaboré et envoyé aux évaluateurs une grille d'évaluation fondée sur la typologie de Rinck (2022), en fournissant pour chaque grand type d'erreur des exemples attestés. Les évaluateurs avaient pour tâche de reporter dans cette grille les erreurs repérées dans chaque texte, selon la typologie proposée (voir figure 1 ci-dessous pour un aperçu).

Nous avons repris la même typologie pour créer des étiquettes dans le logiciel Inception : il s'agit alors, pour les évaluateurs, de sélectionner l'étiquette correspondant à l'erreur repérée (voir figure 2 pour un exemple de ce que l'outil permet).

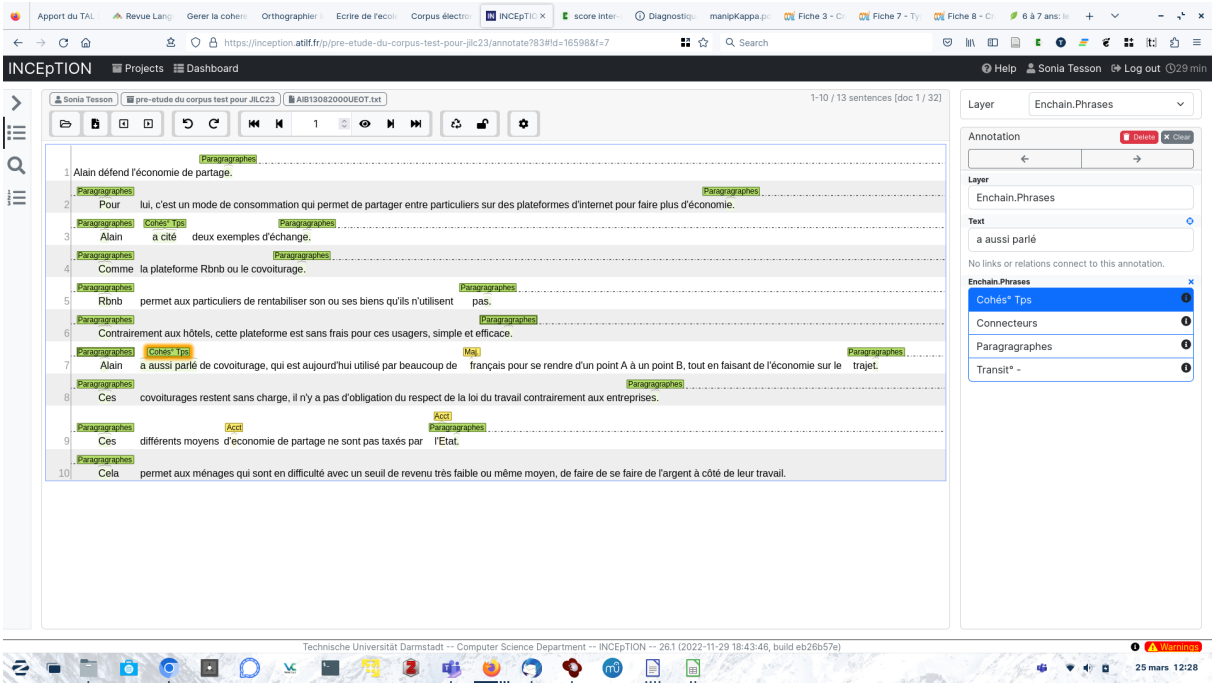
Les éléments de langue devant faire l'objet d'un relevé de la part des évaluateurs sont les suivants :

- Orthographe lexicale ;
- Orthographe grammaticale ;
- Syntaxe ;
- Lexique ;
- Ponctuation ;
- Cohérence.



Aperçu de la grille d'évaluation utilisée pour la recherche

A chaque début d'évaluation d'un texte, puis à la fin du processus, chaque évaluateur est invité à noter l'heure de début et l'heure de fin de son évaluation.



Exemple d'utilisation d'Inception pour catégoriser les erreurs repérées.

A chaque évaluateur, on fait parvenir dans un premier temps le corpus qu'il doit évaluer sans Inception, uniquement à l'aide du tableur (voir figure 1). Une fois cette première phase achevée, on lui fait parvenir le corpus qu'il doit cette fois évaluer au moyen d'Inception, en lui fournissant des informations pour lui permettre de prendre en main l'outil.
Aucun temps de correction n'est imposé aux évaluateurs ; on leur indique seulement que la contrainte est d'essayer d'avoir un degré d'exigence à peu près égal pour toutes les copies.

Méthode d'analyse

Pour chaque méthode et pour chaque corpus, on calcule les éléments suivants :

- Le nombre total d'erreurs repérées (E1 à Ex) ;
- Le nombre total d'erreurs isolées, c'est-à-dire repérées par moins de la moitié des évaluateurs (EI1 à EIx) ;
- Le nombre total d'erreurs communes, c'est-à-dire repérées par plus de la moitié des évaluateurs (EC1 à ECx) ;
- Le taux d'accord inter-annotateurs (AIA) : on calcule la part d'erreurs communes proportionnellement au nombre total d'erreurs repérées (AIA = ECmax x 100 / Emax) ;
- La moyenne du temps passé à l'évaluation par chaque évaluateur.

On cherche également à voir si une ou des catégories d'erreurs sont plus ou moins représentées parmi les EI et les EC.

Une fois ces données obtenues pour chaque méthode, on compare les résultats. L'outil d'annotation de corpus Inception est considéré comme un apport pour une évaluation fiable de productions textuelles :

1. S'il n'est pas trop coûteux en temps : on considère que l'on doit passer sensiblement autant de temps, ou moins avec l'outil que sans lui pour que celui-ci représente un gain pour l'évaluation ;

et si

2. L'AIA est significativement supérieur à celui de l'évaluation sans outil.

Résultats

À l'heure de l'envoi du résumé, l'analyse des résultats n'est pas encore terminée. Nous partagerons les résultats complets pendant les Journées elles-mêmes.

Références bibliographiques

Arrêté du 30 juillet 2018 relatif au diplôme national de licence. Consulté 1 février 2023, à l'adresse https://www.legifrance.gouv.fr/loda/id/LEGIARTI000037296311/2019-09-01/

Bautier, E., & Rayou, P. (2009). *Les inégalités d'apprentissage : Programmes, pratiques et malentendus scolaires* (1re éd). Presses univresitaires de France. https://www-cairn-info.ezproxy.univ-orleans.fr/les-inegalites-d-apprentissage--9782130575276.htm?contenu=sommaire

Boch, F., Sorba, J., & Bessonneau, P. (2016). Évaluer les compétences rédactionnelles : Que tester ? *Le français aujourd'hui*, *N° 193*(2), 127‑144.

Charney, D. (1984). *The Validity of Using Holistic Scoring to Evaluate Writing : A Critical Overview*. 18.

Delarue-Breton, C., & Bautier, É. (2019). Littératie scolaire : Ambitions exigeantes, difficultés de mise en œuvre. *Pratiques. Linguistique, littérature, didactique*, *183‑184*, Art. 183‑184. https://doi.org/10.4000/pratiques.7011

Doquet, C., & Ponton, C. (2021). Écrire de l'école à l'université : Corpus, traitements, analyses outillées. Présentation. *Langue française*, *211*(3), 11‑20. https://doi.org/10.3917/lf.211.0011

Elalouf, M.-L. (2011). Constitution de corpus scolaires et universitaires : Vers un changement d'échelle ? *Pratiques. Linguistique, littérature, didactique*, *149‑150*, Art. 149‑150. https://doi.org/10.4000/pratiques.1702

Morinet, C. (2012). *Du parlé à l'écrit dans les études : Approche théorique et méthodologique de l'articulation entre les pratiques orales et écrites dans l'apprentissage de l'argumentation*. l'Harmattan, DL 2012.

Olive, T., & Piolat, A. (2003). Activation des processus rédactionnels et qualité des textes. *Le Langage et l'Homme*, *28*, 191‑206.

Rinck, F. (2022). *Une approche linguistique dans le champ des littéracies universitaires et avancées* [Thèse d'HDR]. Université Sorbonne nouvelle - Paris 3.

Rot, G., Borzeix, A., & Demazière, D. (2014). Ce que les écrits font au travail. *Sociologie du travail*, *56*(1), Art. 1. https://doi.org/10.4000/sdt.4711

Suchaut, B. (2008). *La loterie des notes au bac : Un réexamen de l'arbitraire de la notation des élèves*. p.19.

Vaguer, C. (2007). Corpus vous avez dit corpus ! De la notion corpus à la création d'un « corpus informatisé ». *Apprendre à fabriquer des corpus à|pour l'école (en chantier)* -, *9*, 207‑223.

La constitution d'un corpus plurisémiotique pour la formation continue et la recherche dans le l'éducation de la petite enfance : une trajectoire collaborative

Anna Claudia Ticca [1,2], Marianne Zogmal [1]
[1] Interaction & Formation, Université de Genève
[2] Laboratoire ICAR, CNRS, Université de Lyon 2
Anna.Ticca@unige.ch, Marianne.Zogmal@unige.ch

Introduction

Cette contribution présente la démarche entreprise dans la constitution d'un corpus plurisémiotique lors de la réalisation d'un dispositif de formation continue, qui mobilise l'analyse de l'activité dans une perspective interactionnelle. Cette formation, répartie sur deux ans, est mise en place dans le champ professionnel de l'éducation de l'enfance en Suisse. Les métiers de la prise en charge d'autrui présentent des spécificités ; l'une d'elles est constituée par la dimension interactionnelle et collective des activités, entre enfants, collègues, apprenants ; une autre réside dans le rôle central de l'observation fine des situations, permettant d'ajuster les modalités des actions éducatives. Dans leur travail quotidien, les professionnel.le.s observent les conduites des enfants et portent ainsi un certain regard analytique sur ce qui se passe dans le déroulement interactionnel, de façon située. Lors des séances de formation, les participantes s'engagent à réaliser des films, à sélectionner des extraits et à effectuer des transcriptions. Elles constituent ainsi un corpus de données portant sur des situations d'interactions au travail. Le recueil et l'analyse de données empiriques représentent des outils sémiotiques qui permettent de soutenir la réflexivité et les capacités d'observation sur les pratiques professionnelles. En effet, ce dispositif de formation est conçu à partir de l'hypothèse que des pratiques mises en place dans le contexte de la recherche peuvent constituer des ressources pour la formation. Pour étudier les pratiques de constitution et de mobilisation d'un corpus plurisémiotique en formation, une perspective interactionnelle en analyse des activités de travail offre des outils théoriques et méthodologiques pertinents (Filliettaz & al, 2021).

Corpus et méthodologie

Corpus

Le corpus plurisémiotique constitué dans le cadre de ce dispositif de formation continue est réparti dans deux volets proposés au même groupe de participantes. Lors de la première année de formation (volet I), 14 éducatrices ont participé à une formation d'une durée d'environ quatre journées au total. La deuxième année de formation (volet II) a été suivie par 11 éducatrices (env. trois journées au total). Les séances de formation sont filmées à des fins de

recherche et permettent de recueillir, au total, des données vidéo d'une durée d'environ 25 heures pour le volet I et d'environ 23 heures pour le volet 2.

Sur le plan procédural, la constitution du corpus plurisémiotique s'accomplit par différentes phases d'activités successives. Pour étudier ce processus dynamique et collectif, nous nous intéressons plus spécifiquement à la constitution d'une partie du corpus linguistique portant sur l'activité d'une des participantes. Pendant la première année de la formation, chacune des éducatrices du groupe identifie une situation à observer dans le cadre de son travail et réalise (a) l'enregistrement d'une situation d'interaction entre adultes et enfants. Elle repère également (b) un court extrait dans le film réalisé et en produit (c) une transcription. La séquence filmée est choisie en fonction des intérêts personnels et elle est ensuite soumise pour l'analyse à l'ensemble du collectif. Cette séance d'analyse collective est filmée (e) et retranscrite (f) par l'équipe de chercheur.es / formateur.es pour des fins de recherche.

Dans la suite, certaines éducatrices ayant participé à la formation décident de réutiliser le film crée précédemment dans le cadre d'une activité collective au travail. Ceci concerne notamment une des participantes qui choisit de présenter et analyser l'extrait du film avec des collègues lors d'une rencontre désignée comme 'colloque pédagogique'. Cette situation d'analyse (II.a, voir le tableau ci-dessus), organisée et animée de façon autonome, est également filmée. Ce film portant sur le colloque d'équipe fera l'objet de la séance de formation du Volet II afin de mieux comprendre les stratégies d'analyse et d'animation de l'éducatrice concernée. La table 1 montre les constituants du corpus réalisés par cette participante individuellement (Volet I (a), (b), (c), (d)/ Volet II (a), (b), (c), (d)) ainsi que les constituants du corpus portant sur la participante concernée (Volet I (e), (f)/Volet II (e), (f)) :

Temporalité	Données constituées dans le processus de formation par une des participantes				Données constituées dans le processus de recherche portant sur la participante concernée	
Volet I	(a) Film de référence (situation de travail)	(b) Extrait du film de référence	(c) Transcription du film de référence	(d) Présentation de synthèse	(e) Film portant sur la séance d'analyse	(f) Transcription portant sur la séance d'analyse
Volet II	(a) Film de référence (situation d'analyse dans le contexte du travail)	(b) Extrait du film de référence	(c) Transcription du film de référence	(d) Présentation de synthèse	(e) Film portant sur la séance d'analyse	(f) Transcription portant sur la séance d'analyse

table 16. : **Constituants du corpus portant sur l'activité d'une des participantes à la formation**

L'ensemble des données empiriques recueillies est ensuite importé sur une base de données du logiciel *Transana Multiuser*. Le repérage des séquences qui montrent la mobilisation des données du corpus linguistique (films vidéo, transcriptions) est réalisé à l'aide du logiciel de *MindManager.* Les données sont ensuite analysées selon les principes de l'approche interactionnelle (Filliettaz, 2018).

Méthodologie

Le dispositif de formation présenté dans cette contribution s'inscrit dans un programme de formation qui vise au développement d'une posture analytique basée sur l'observation et description des phénomènes interactionnels tels qu'ils se déploient dans l'activité éducative. Cependant, cet enrichissement et cette transformation des postures analytiques ne s'opèrent pas à n'importe quelles conditions. Afin de soutenir la co-construction d'un regard analytique, une démarche de recherche-intervention met en œuvre un processus participatif. Le dispositif mis en place s'inscrit ainsi dans une continuité avec des travaux portant sur des démarches collaboratives (Desgagné, 1997 ; Snow, 2004 ; Veillard & al., 2011). L'expérimentation de la démarche d'analyse interactionnelle par les participantes elles-mêmes constitue un élément central de cette démarche participative. Ce sont les éducatrices mêmes qui contribuent à la constitution d'une partie du corpus plurisémiotique utilisé en formation et pour des fins de recherches. Une telle approche vise à favoriser une posture analytique qui permet aux éducatrices de s'intéresser aux points de vue endogènes des participants à une interaction, en considérant que les « comportements interactionnels, dans leurs dimensions à la fois verbales, paraverbales et non-verbales, rendent manifestes des éléments de signification que les participants eux-mêmes attribuent à leurs conduites et à celles de leurs partenaires » (Filliettaz, 2018, pp. 53-54). Les analyses que nous menons (Ticca & al., à paraître) illustrent comment les participantes à la formation construisent progressivement une forme interprétée et intelligible de l'action observée dans le film, et comment la constitution du corpus plurisémiotique y contribue. Afin d'illustrer l'utilisation de corpus constitué de manière collaborative nous nous focalisons sur le parcours au sein du dispositif d'une éducatrice dans les deux moments de la formation (volet I et volet II). Nous présenterons quelques extraits des films en situation de formation, dans lesquels l'éducatrice endosse le rôle d'animatrice, ainsi que lors du colloque pédagogique, dans lequel l'éducatrice gère la séance de manière autonome. Dans les deux situations, le film de référence utilisé est le même (une situation d'interaction au travail). Nous montrerons également comment ce corpus plurisémiotique, constituée par la participante, s'inscrit dans l'ensemble du corpus de données recueillies au sein du dispositif de formation.

Résultats

Les différentes ressources matérielles créées par les éducatrices, et qui constituent une partie du corpus plurisémiotique, offrent des opportunités de formation variées :

- La réalisation du film vidéo dans le contexte du travail (volet I) ou lors d'un colloque d'équipe (volet II) amène chaque éducatrice à identifier des situations qui lui permettent d'aborder un questionnement d'ordre professionnel. L'enregistrement vidéo favorise l'observation des pratiques professionnelles pour repérer ce qui est montrable ou non dans un contexte de formation et pour recueillir une trace observable de l'activité de travail auprès des enfants ou des collègues.
- La sélection de l'extrait du film implique d'élaborer un premier questionnement analytique ; dans cette perspective, l'extrait est utilisé comme une ressource épistémique permettant d'accéder à des contenus ciblés.
- Lors de la séance d'analyse, l'outil vidéo prend des formes et significations diverses. Il constitue une trace des éléments observables dans l'activité de travail. Il est aussi utilisé comme un outil d'animation et d'analyse, permettant à l'animatrice à la fois de

montrer des contenus précis, selon son agenda, et de laisser découvrir aux participantes elles-mêmes les phénomènes pertinents.

- La transcription de l'extrait choisi produite par les éducatrices est mobilisée en situation de formation en tant qu'objet multimodal et épistémique. La transcription permet en effet de reconstruire le déroulement temporel de la situation de référence dans sa dimension praxéologique et multimodale. Gestes, regards, manipulation d'objets mais aussi silences, pauses, hésitations, énoncés deviennent des outils analytiques permettant l'élaboration d'un argument ou d'un point de vue analytique. Les transcriptions peuvent être mobilisées pour (contre)argumenter un point de vue, réorienter l'analyse, promouvoir une transition topicale, etc.

Le recueil des données filmées sur l'ensemble des séances de formation permet ensuite de repérer l'organisation collective des activités lors des démarches de formation et les délimitations des étapes de leur déroulement. Pour mieux étudier le processus participatif aboutissant à la constitution d'un corpus plurisémiotique à partir des choix effectués par les éducatrices, nous analysons la manière dont l'utilisation et l'exploitation des outils sémiotiques à disposition évolue au fil des séances. Observer l'utilisation et la mobilisation du corpus vidéo et de la transcription en tant qu'objets sémiotiques et épistémiques représente un moyen inédit d'accéder aux différents types de savoirs qui se développent et s'imbriquent lors des séances de formation. L'intérêt pour une telle mobilisation du corpus linguistique vise à mieux comprendre les effets d'une démarche participative lors de la constitution d'un corpus plurisémiotique dans un dispositif de formation continue.

Références bibliographiques

Desgagné, S. (1997). Le concept de recherche collaborative : l'idée d'un rapprochement entre chercheurs universitaires et praticiens enseignants. Revue des sciences de l'éducation, XXIII(2), 371-393.

Filliettaz, L. (2018). *Interactions verbales et recherche en éducation : principes, méthodes et outils d'analyse.* Université de Genève, Carnets des sciences de l'éducation.

Filliettaz, L., Bimonte, A., Koleï, G., Nguyen, A., Roux-Mermoud, A., Royer, S., Trébert, D., Tress, C., & Zogmal, M. (2021). Interactions verbales et formation des adultes. *Savoirs, 56(2),* 11-51.

Ticca A.C., Zogmal, M. & Filliettaz, L. (à paraître). Former des éducateurs à l'analyse des interactions : quelle place au langage ? Dans M.-A.. Akinci, I. Maillochon, V. Miguel Addisu (dirs.). *Vivre et parler avec le jeune enfant en crèches multi-accueil*. L'Harmattan.

Snow, C.E. (2004). Rigor and realism: Doing educational science in the real world. *Educational Researcher, 44(9),* 460-466.

Veillard, L., Tiberghien, A. et Vince, J. (2011). Analyse d'une activité de conception collaborative de ressources pour l'enseignement de la physique et la formation des professeurs : le rôle de théories ou outils spécifiques. *Activités, 8(2),* 202-227.

The role of crime and verdict in the defendants' and victims' use of degree adverbs in the Late Modern English courtroom discourse

Aditya Upadhyaya [1] et Billie Anjellyn Craig [1]

[1] Department of English, University of Giessen (JLU) Aditya.Upadhyaya-2@anglistik.uni-giessen.de, Billie.A.Craig-2@anglistik.uni-giessen.de

Introduction

Human beings are emotional creatures. Words are the conveyor of concepts and meanings for us. In this light, degree adverbs like *very*, *so* and *truly*, are typically the modifying adverbs used as a scaling device (Quirk et al. 1985 : 445) to indicate the commitment and certainty of the speaker to the "truth value of the proposition" (Simon-Vandenbergen 2008 : 1521). Given the association of degree adverbs with emotional language (Tagliamonte and Roberts 2005 : 290) and in building a speaker's image, their use in a courtroom is an interesting research domain. Victims and defendants need to boost the credibility of their statements, as the courtroom discourse is likely to result in momentous decisions in their lives. Moreover, Svongoro et al. indicate that in the courtroom "the key components are interpretation and persuasion, and the magistrate is the ultimate interpreter of all messages, and has to be persuaded, largely on the basis of linguistic evidence" (2012 : 125). In this context, it will be interesting to study how defendants or victims who lack formal training in special language use, and whose mental, social, and physical circumstances might influence their argument/expression, use degree adverbs to emphasise their points in courtrooms. Despite a vast body of research on degree adverbs, a deductive linguistic analysis of the feature in legal proceedings, a suitable resource for studying language behaviour, is still lacking. While some studies focused on the influence of factors like social class, gender, and time on the use of degree adverbs in the courtroom (Claridge et al. 2020 : 866, Claridge et al. 2021 : 68), the studies usually ignore potentially significant factors like offence category and verdict. These two factors may affect the use of degree adverbs in accounts of untutored speakers in a highly regimented legal environment. The present study takes a forensic linguistics approach and focuses on the frequency development of the degree adverbs in the defendants' and the victims' discourse in the proceedings of the Old Bailey, the criminal court of London from 1720–1913, using the Old Bailey Corpus (the OBC) (Huber et al. 2012). The variables investigated are : OFFENCE (theft, other non physical crimes, and physical crimes), VERDICT (guilty, not guilty), SOCIAL CLASS (higher, lower), TIME (18^{th}, 19^{th}, 20^{th} centuries), GENDER (male, female), SPEAKER ROLE (victim, defendant). Degree adverbs (I) modify verbs, adjectives, adverbs, prepositions, (II) intensify the modified constituent and in turn, modify its illocutionary force, and (III) reflect the positionality of the speaker (Rhee 2016 : 399 ; see also Quirk et al. 1985, Huddleston and Pullum 2002, Biber et al. 2002, Bordet 2017). The following adverbs were chosen for the present study : *very*, *so*, *entirely*,

absolutely, *certainly*, *at all*, *extremely*, *fully*, *hardly*, *pretty*, *really*, *totally*, and *truly*. They are also the most frequent degree adverbs in the defendants' and victims' speech in the corpus. To explore the significance of degree adverbs in the Late Modern English courtroom language, the present study seeks to explore the following facets :

- The frequency development of the degree adverbs in the defendants' and victims' speech in the OBC from 1720-1913
- The effect of crime, verdict and the speakers' role in the courtroom on the frequency of use of the degree adverbs
- The effect of the speakers' gender and social class in their use of the degree adverbs.

Hypotheses

Using findings of prior research as a guide, the following hypotheses have been tested in the following experiments :

H1 : The speakers associated with property crimes like theft are anticipated to use degree adverbs more often than those associated with other crimes. In regard to defendants, this can be due to the fact that most capital punishments were administered for property crimes as an idea of justice (Emsley et al. 2018 : n.pag.). As for the victims, it may have been due to the fact that the concept of wealth and property was of paramount importance to people, particularly in the 18^{th} century (Hay 1975 : 19).

H2 : The defendants of the 18^{th} century are likely to use more degree adverbs than those of the 19^{th} and 20^{th} centuries since the 'bloody code' of the 18^{th} century imposed the death penalty even for trivial crimes (Hay 1975 : 20-22) and they also operated "under severe disadvantages" (Emsley et al. 2018 : n.pag.). In addition, victims of the 20^{th} century are likely to use more degree adverbs than their counterparts in the preceding two centuries. This may be due to the change in legal practices in the latter period, which lessened the role of victims as prosecutors (Hitchcock and Turkel 2016), resulting in the use of more degree adverbs to defend their argument.

H3 : The speakers associated with 'not guilty' verdict are expected to use degree adverbs more than those associated with 'guilty' to enhance the illocutionary force of the speech act expressing high certainty or conviction concerning its validity.

H4 : It is expected that the working class speakers of the 18^{th} century will use more degree adverbs than their counterparts from higher social classes as the law was geared toward the rich and those in higher social positions (Hay 1975 : 44-45).

H5 : The female speakers in the courtroom are expected to use more degree adverbs than the males as degree adverbs are typically owned by women (Tagliamonte and Roberts 2005 : 284, Lakoff 1975).

Corpus and Methodology

The data for the present study comes from the Old Bailey Corpus 3.0 (https://obc-client.de/index.html), which is a textually, pragmatically, and sociolinguistically annotated corpus containing 637 published trial proceedings of the Old Bailey that took place from 1720 to 1913 (Huber et al. 2016 : 20-21).

	Female	Male	Higher Class	Lower Class
Defendant	312,452	1,346,380	224,006	373,292
Victim	774,857	3,435,691	1,418,328	1,265,798

table 1. : Word counts of the defendants and victims in the OBC 3.0 between 1720-1913

Table 1 depicts the overall disproportion in gender, which is due to the fact that the proportion of women offenders tried at the Old Bailey was low. Similarly, working-class defendants dominated more than the higher-class defendants as most of the lawbreakers came from society's poorer sections (Emsley 2005 [1987] : 57, Oberwittler 1990 : 5). All the instances of the chosen degree adverbs in the OBC were retrieved using the annotated OBC 3.0 concordancer, which generates Keyword-in-Context along with the metadata. The raw data were screened to retrieve all the instances of the relevant construction employed by the defendants and victims only. Every concordance was then meticulously analysed to get rid of the occurrences irrelevant to the research. A particularly challenging factor of the data is that the metadata in the OBC 3.0 does not provide the total word count of every speaker in a single row, but identical speakers are listed separately according to different turns per speaker. However, this complexity was dealt with using Python 3.7 (Van Rossum et al. 2009), which is a powerful programming tool to solve any task such as transforming raw dataset into a comprehensible format with the help of organised syntax (Agarwal 2015 : 30). A number of python algorithms were executed to add up (1) the total number of degree adverbs spoken by the speakers (2) the total word count of every speaker in a trial with identical IDs. After data cleansing, the search yielded a total of 1735 (by 918 defendants) and 4881 (by 2419 victims) relevant hits of the 13-degree adverbs. Finally, the frequency of degree adverbs spoken by a speaker was normalised to a common base of a hundred thousand words (phtw). For the next step, multifactorial test, multiple linear regression was performed on the data to predict the effect of the predictors on the response variable (normalised frequency of degree adverbs). According to Gries, the main aim of regression modelling is to determine "how much one can 'predict' what a response does depending on what one or more predictors do" (2021 : 238). Thus, multiple linear regression is an ideal test to evaluate the relationship between the predictors and the response variable by assessing the "unique contribution of each individual parameter (predictor), holding the other parameters (predictors) constant" (Claridge et al. 2021 : 77).

Résultats

An initial exploratory (descriptive) analysis revealed the skewness of the dependent variable (fre- quency of degree adverbs) and thus, a log-transformed version of the dependent variable was adopted for the statistical modelling. The final model demonstrates (Figure 1) that a) crime (*p value* = 3.702e- 07 ***), b) interaction between speaker role and time (*p value* < 2.2e-16 ***), c) interaction between speaker role and verdict (*p value* = 0.01182 *), d) interaction between social class and time (*p value* = 0.00274 **) and e) interaction between verdict and time (*p value* = 0.01507 *) significantly influenced the number of degree adverbs in the speakers' speech.

- As to the speakers' association with crime and their use of degree adverbs, Figure 1 (a) discloses that the speakers associated with theft significantly spoke the highest number of degree adverbs (10.14 phtw) followed by those associated with physical crimes (10 phtw) whereas those connected with other non physical crimes (coded as ONP) used the least (9.8 phtw). These results are consistent with *H1*.

- Notably, the defendants of the 18^{th} and 19^{th} centuries dominated in their use of degree adverbs with c. 10.7 and 10.6 instances phtw respectively (Figure 1 (c)). However, in the 20th century, it was the victims who spoke more degree adverbs (9.3 phtw) than the defendants (8.2 phtw). These differences in the speakers' use of degree adverbs across the three centuries are highly significant. These results align with *H2*.

- As determined by speaker role across different verdicts, no significant differences were found in the use of degree adverbs in defendants' speech. In contrast, victims who were disfavored by the court (i.e., who could not prove their alleged offenders 'guilty') used significantly more degree adverbs (10.00 phtw) than those who were not (9.8 phtw) (Figure 1 (b)). It is worth noting that in general, speakers associated with the 'not guilty' verdict used the greatest number of degree adverbs in the 18^{th} century (10.53 phtw), whereas the speakers associated with the 'guilty' verdict used the least of them in the 20^{th} century (8.95 phtw) (Figure 1 (d)). The results partially support *H3*.

- Moreover, degree adverbs were significantly more dominant among working class speakers of the 18^{th} century (10.54 phtw) than their 19^{th} (9.95 phtw) and 20^{th} century (9.06 phtw) counterparts, while higher-class speakers of the 20^{th} century (8.95 phtw) used them the least (Figure 1 (e)). H4 is supported.

- No significant difference emerged in the use of degree adverbs by men and women, which paints a complex picture of language use, discourse mode and speaker role. In other words, gender does not always predict language use, rather language use is more closely affiliated with contextual factors like the speaker-listener relationship (Janssen and Murachver 2004 : 349). Likewise, men and women defendants and victims, lacking social power (Hildebrand-Edgar and Ehrlich 2017 : 92) in the highly institutionalised courtroom, seem to display a similar pattern in their use of degree adverbs for "message intensity" (see McEwen and Greenberg 1970) and override the gender effect to meet the requirements of their role and situation.

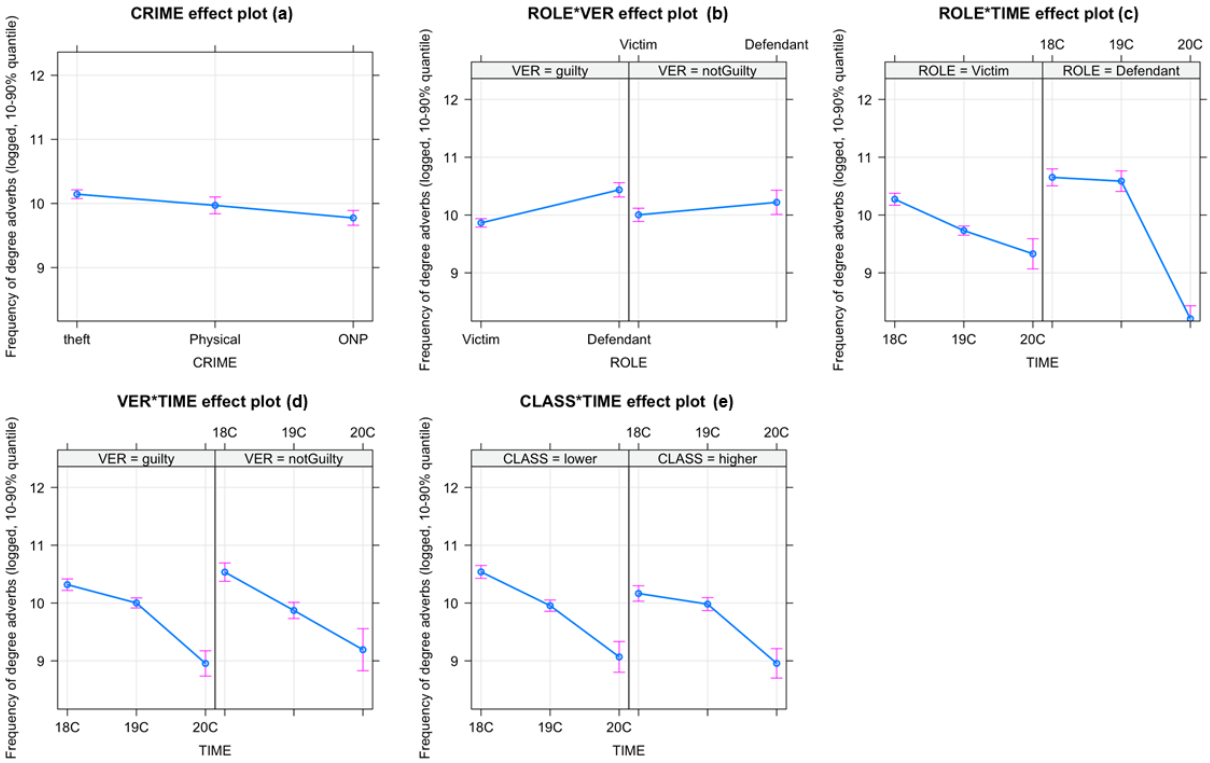


figure . 1 Frequency of degree adverbs in the speakers' discourse across the significant predictors

Discussion

While previous research has so far paid attention to degree adverbs with respect to collocations (Simon-Vandenbergen 2008 ; Bäcklund 1976 ; Ito and Tagliamonte 2003), semantic and structural characteristics (Jespersen 1922 ; Bolinger 1972 ; Stoffel 1901), sociolinguistic factors in the courtroom (O'Barr and Conley 1973 ; O'Barr 1982 ; Claridge et al. 2020), the current study opens up an interesting line of discussion in forensic linguistics by taking crime and verdict into account as these variables have the potential to influence the psychological profile of an individual and the language of incidence. Crucially, none of the two variables have been investigated hitherto from a socio-linguistics/forensics standpoint using a larger database like the OBC.

With regard to social class, criminality was seen as a class problem essentially in the 18th and 19th century when most of the delinquents presented before the courts came from the lower stratum of society. While the criminal law's nature gave huge discretion to rich men except for the prosecutor and because the law did not permit the offenders to hire an attorney for addressing the jury, the poor's defence was frequently "a halting, confused statement" (Hay 1975 : 42). In addition, the rhetoric of degree adverbs is associated with emotional language and is a response to argumentative threat especially when the defendants (and victims) were on the losing side (Long and Christensen 2012 : 959), which explains the predominant use of degree adverbs by the working class.

Moreover, it can be assumed that the lower use of degree adverbs by the victims might have been a part of a carefully rehearsed 'powerful' language strategy. On the contrary, the change in the legal practices in the later period weakened the role of victim as a prosecutor (Hitchcock and Turkel 2016 : 946), which may have been one of the reasons of a greater use of degree adverbs by the victims in the 20th century. However, such an interrelatedness is just

an assumption and there may be other variables such as the speaker's mental and inner state that contributed to differences in the number of degree adverbs in the speakers' accounts.

Since crime turned out to be a significant predictor of the distribution of degree adverbs, it can be considered a potential factor in present-day English (PDE) and legal context in predicting such a distribution, especially in contemporary defendants' narratives who are at a much advantageous position with a defence attorney at their disposal and presumed "to be innocent until proven guilty" unlike their historical equivalents (Jucker 2012 : 205-206). The legal implications of the differences in the use of degree adverbs in response to an allegation of different crimes cannot be undervalued.

Centred on the above findings, it would be beneficial to have a similar corpus like the OBC based on the proceedings of modern courtroom trials that would allow linguists (chiefly forensic linguists) to characterise the role of degree adverbs in boosting the pragmatic accuracy of interlocutors across sociopragmatic and legal factors. This study may be considered a little fragment in adding to the understanding of the use of degree adverbs in Late Modern English, which has attracted sizeable attention recently.

Bibliography

Agarwal, Vivek (2015). "Research on Data Preprocessing and Categorization Technique for Smartphone Review Ana- lysis." International Journal of Computer Applications 131.4 : 30-36. Web.

Bäcklund, Ulf (1973). The collocation of adverbs of degree in English. Uppsala : Almqvist and Wiksell. Web.

Biber, Douglas, Conrad, Susan, and Leech, Geoffrey (2002). Longman Student Grammar of Spoken and Written English. England : Pearson Education Limited. Web.

Bolinger, Dwight (1972). Degree Words. Boston : De Gruyter Mouton. Web.

Bordet, Lucile (2017). "From vogue words to lexicalized intensifying words : the renewal and recycling of intensifiers in English. A case-study of very, really, so and totally." Lexis 10 : 1-16. Web.

Claridge, Claudia, Jonsson, Ewa and Kytö, Merja (2020). "Entirely innocent : a historical sociopragmatic analysis of maximizers in the Old Bailey Corpus." English Language and Linguistics 24.4 : 855-874. Web.

Claridge, Claudia, Jonsson, Ewa and Kytö, Merja (2021). "A Little Something Goes a Long Way : Little in the Old Bailey Corpus." English Language and Linguistics 49.1 : 61-89. Web.

Emsley, Clive (2005). Crime and Society in England : 1750-1900 [1987]. UK : Pearson Education Limited. Web.

Emsley, Clive, Hitchcock, Tim, and Shoemaker, Robert (2018). "Trial Procedures." Old Bailey Proceedings Online. Web. <https ://www.oldbaileyonline.org/static/Trial-procedures.jsp> version 7.0 (2 January 2023).

Gries, Stefan Th. (2021). Statistics for Linguistics with R : A Practical Introduction. Berlin/Boston : De Gruyter Mouton. Web.

Hay, Douglas (1975). “Property, Authority and the Criminal Law.” Albion’s Fatal Tree : Crime and Society in Eighteenth- century England. Ed. Douglas Hay, Peter Linebaugh, John G. Rule, E.P. Thompson, and Cal Winslow. New York : Pantheon Books. 17-64. Web.

Hay, Douglas (1980). “Crime and justice in eighteenth-and nineteenth-century England.” Crime and Justice 2 : 45-84. Web.

Hildebrand-Edgar, Nicole, and Ehrlich, Susan (2017). “She was quite capable of asserting herself” : Powerful Speech Styles and Assessments of Credibility in a Sexual Assault Trial.” Language and Law 4.2 : 89-107. Web.

Hitchcock, Tim, and Turkel, William J. (2016). “The "Old Bailey Proceedings, 1674 – 1913" : Text Mining for Evidence of Court Behavior.” Law and History Review 34.4 : 929-955. Web.

Huber, Magnus, Magnus Nissel, Patrick Maiwald, and Bianca Widlitzki (2012). The Old Bailey Corpus (OBC). 1720- 1913. Web. <https ://obc-client.de/index.html> (1 January 2023).

Huber, Magnus Nissel, Karin Puga (2016). Old Bailey Corpus 2.0. Web. <hdl :11858/00-246C-0000-0023-8CFB-2> (2 January 2023).

Huddleston, Rodney, and Pullum, Geoffrey K. (2002). The Cambridge Grammar of the English Language. England : CUP. Web.

Ito, Rika, and Tagliamonte, Sali (2003). “Well weird, right dodgy, very strange, really cool : Layering and recycling in English intensifiers.” Language in Society 32 : 257-279. Web.

Janssen, Anna, and Murachver, Tamar (2004). “The relationship between gender and topic in gender-preferential language use.” Written Communication 21.4 : 344-367. Web.

Jespersen, Otto H. (1922). Language : Its nature, development, and origin. London : George Allen and Unwin. Web. Jucker, Andreas H. (2012). “Pragmatics and discourse.” English Historical Linguistics : An International Handbook. Ed. Alexander Bergs, and Laurel J. Brinton. Germany : De Gruyter Mouton. 197-211. Web.

Lakoff, Robin (1975). Language and Woman’s Place. New York : Harper and Row. Web.

Long, Lance N., and Christensen, William F. (2012). “When Justices (Subconsciously) Attack : The Theory of Argu- mentative Threat and the Supreme Court.” OR. L. Rev. 91 : 933-960. Web.

McEwen, William J., and Greenberg, Bradley S. (1970). “The Effects of Message Intensity on Receiver Evaluations of Source, Message and Topic.” The Journal of Communication 20 : 340-350.

O’Barr, William M., and Conley, John M. (1976). “When a Juror Watches a Lawyer.” Barrister 3 : 42-45. Web. O’Barr, William M. (1982). Lingustic Evidence : Language, Power and Strategy in the Courtroom. California : Academic Press Limited. Web.

Oberwittler, Dietrich (1990). “Crime and Authority in Eighteenth Century England : Law Enforcement on the Local Level.” Historical Social Research / Historische Sozialforschung 15.2 : 3-34. Web.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey, Svartvik, Jan (1985). A Comprehensive Grammar of the English Language. New York : Longman. Web.

R Core Team (2020). R : A language and environment for statistical computing : R Foundation for Statistical Computing. Vienna. Web. < https ://www.R-project.org/> (2 January 2023).

Rhee, Seongha (2016). “On the emergence of the stance-marking function of English adverbs : A case of intensifiers.” Linguistic Research 33.3 : 395-436. Web.

Simon-Vandenbergen, Anne-Marie (2008). “Almost certainly and most definitely : Degree modifiers and epistemic stance.” Journal of Pragmatics 40 : 1521-1542. Web.

Stoffel, Cornelis (1901). Intensives and Downtoners : A study in English adverbs. Heidelberg : Winter’s universitätsbu- chhandlung. Web.

Svongoro, Paul, Mutangadura, Josephat, Gonzo, Lameck, and Mavunga, George (2012). "Language and the legal pro- cess : A linguistic analysis of courtroom discourse involving selected cases of alleged rape in Mutare, Zimbabwe." South African Journal of African Languages 32.2 : 117-128. Web.

Tagliamonte, Sali, and Roberts, Chris (2005). "So weird ; so cool ; so innovative : The use of intensifiers in the television series Friends.” American Speech 80.3 : 280-300. Web.

Van Rossum, G., and Drake, F.L. (2009). Python 3 Reference Manual. Scotts Valley, CA : CreateSpace.

Étude comparative d'éléments du lexique scientifique français/chinois dans une perspective didactique

Rui Yan et Sylvain Hatier
Laboratoire LIDILEM, Université Grenoble Alpes
rui.yan@univ-grenoble-alpes.fr, sylvain.hatier@ univ-grenoble-alpes.fr

Introduction

Le lexique scientifique transdisciplinaire (désormais LST) joue un rôle primordial dans l'écrit scientifique puisqu'il est au cœur de l'argumentation et de la structuration du discours scientifique (Drouin, 2007; Paquot & Bestgen, 2009). Il est notamment étudié dans une perspective didactique, que ce soit pour la création de ressources didactiques (Coxhead, 2002; Paquot, 2010), de manuels scolaires (Phal & Beis, 1972) ou dans le cadre de l'aide à la rédaction scientifique (Pecman, 2004 ; Tran, 2014 ; Yan et Hatier, 2016).

Dans cette étude, nous souhaitons mener une analyse lexicale comparative parmi un échantillon de verbes et de noms du LST en français et en chinois. L'intérêt de ce travail se justifie premièrement d'un point de vue didactique par le nombre important d'étudiants sinophones inscrits dans les universités françaises, ce public étant par ailleurs souvent confronté à des difficultés rédactionnelles (Cavalla, 2018). D'autre part, bien que certains auteurs (Hu et Gao, 2011 ; Loi et Lim, 2013 ; Mu et al., 2015) s'intéressent au discours scientifique chinois dans une perspective contrastive et didactique, il manque à notre connaissance de travaux effectuant une description linguistique du lexique spécialisé à l'œuvre dans ce genre d'écrit.

En ce qui concerne le travail d'identification du LST en chinois, les travaux de Liu et al. (2016) ont permis de constituer une *Chinese Academic Wordlist* basée sur la fréquence et la dispersion dans un corpus de 1000 articles scientifiques (Chen et Tao, 2019). Nous pouvons également mentionner qu'un travail de traduction en chinois des entrées françaises du LST est en cours sur la plateforme lexicale issue des travaux de (Hatier et al., 2016).

L'intérêt de la linguistique de corpus pour la caractérisation de genre d'écrit, tel l'écrit scientifique à travers son lexique, a ainsi été mis en évidence tant du point de vue de la description linguistique que dans une perspective didactique (Flowerdew, 2005).

Corpus et méthodologie

Corpus

Afin de pouvoir caractériser les éléments du lexique en français et en chinois, nous disposons de deux corpus composés d'articles de recherche dans 4 disciplines des sciences humaines et sociales. Ces articles sont formatés au format XML puis analysés en dépendances afin de pouvoir exploiter les informations de combinatoires, tant au niveau des colligations, des collocations que des structures phraséologiques plus complexes.

Le tableau ci-dessous présente les deux corpus utilisés :

	Corpus français (Termith-SHS[125])	Corpus chinois
Articles	200	240
Tokens	1.76 millions	1.74 millions
Disciplines	Sciences de l'éducation, économie, sociologie, linguistique	Sciences de l'éducation, économie, sociologie, linguistique

Tableau 1: Composition des corpus d'analyse

Méthodologie

L'intégration des corpus d'analyse à l'outil d'exploration de corpus Lexicoscope (Kraif, 2016) nous permet ensuite d'analyser diverses propriétés des entrées du lexique choisies. Au niveau textométrique, l'analyse des fréquences (globales et par disciplines) et de la dispersion permet de s'assurer du caractère transdisciplinaire et récurrent des éléments du LST en chinois. Aux niveaux lexicométrique et phraséologique, l'analyse des cooccurrents les plus significatifs, des arbres lexico-syntaxiques récurrents (Tutin & Kraif, 2016) ainsi qu'une analyse manuelle des concordances nous permet de caractériser les entrées du LST en chinois et de mettre en exergue les points convergents et divergents entre les entrées du LST dans les deux langues.

Les entrées du LST français ont été précédemment élaborées dans le cadre du projet TermITH. Le lexique a été extrait automatiquement puis validé manuellement par comparaison d'un corpus diversifié d'articles en sciences humaines et sociales et d'un corpus de contraste de grande échelle, tous deux analysés en dépendances. Structurée en classes et sous-classes sémantiques, cette ressource est composée de 2 310 entrées, couvrant les catégories des noms, adjectifs, verbes et adverbes (Hatier et al., 2016).

Concernant les éléments du LST en chinois, une liste de mots candidats au statut de LST a été extraite en retenant les critères de fréquence et de répartition, sur le même modèle que pour le LST français[126]. Nous avons ensuite manuellement sélectionné, parmi les plus fréquents (pour que les informations de combinatoire soient les plus riches possibles), les éléments lexicaux correspondant à la définition du LST.

Nous nous concentrons ainsi sur 10 verbes et 10 noms du LST chinois pour explorer le corpus. Comme point de départ de l'exploration du corpus français, nous analysons les propriétés linguistiques des équivalents traductionnels français de ces entrées chinoises.

L'analyse comparative permet de relever certaines caractéristiques propres au LST dans l'écrit scientifique français et chinois, propriétés importantes dans le cadre de l'aide à la rédaction scientifique. Plusieurs études ont ainsi mis en avant l'intérêt didactique de proposer aux apprenants scripteurs ces informations d'ordre lexical, syntaxique et sémantique afin d'optimiser l'appropriation du lexique (Tran & Royer, 2020). Ainsi, les données telles que les

[125] TermITH (Terminologie et Indexation de Textes en sciences Humaines) : ANR-12-CORD-0029 CONTINT : https://web-data.atilf.fr/ressources/termith/index.php

[126] Le critère de spécificité n'a pas encore été intégré, le corpus de contraste pour le chinois n'ayant pas encore été constitué

classes sémantiques (dont nous disposons pour les entrées françaises) se révèlent également une entrée didactique plus abordable (Cavalla, 2019) dans le cadre de l'enseignement/apprentissage d'un lexique spécialisé.

Analyse contrastive et premiers résultats

Dans un premier temps, l'étude de cet échantillon du LST permettra de mettre en évidence les rôles de chercheur assumés par le scripteur (Fløttum et al., 2006) tant dans le corpus français que chinois et d'examiner ainsi certaines facettes de la démarche scientifique (l'établissement du constat scientifique, l'analyse des données ainsi que l'interprétation des résultats).

Nous examinerons par la suite les routines (Tutin & Kraif, 2014) liées à ces verbes et à ces noms afin d'étudier les fonctions rhétoriques associées et propres au discours scientifique dans nos corpus d'analyse. L'analyse comparative de ces routines a ici pour but de mettre en évidence, dans les écrits français, les routines saillantes dont la maîtrise est essentielle pour améliorer la rédaction scientifique des étudiants sinophones. L'intégration de la phraséologie et de sa dimension rhétorique dans l'étude du lexique se justifie ainsi de par son apport didactique reconnu dans l'enseignement du lexique (Cavalla, 2021).

Les premières observations dans les corpus montrent par ailleurs des différences liées aux dimensions énonciative et argumentative. Les premières analyses montrent une tendance plus élevée chez les auteurs francophones à assumer les rôles d'argumentateur et d'évaluateur dans l'écrit scientifique par rapport aux auteurs chinois qui, selon notre hypothèse, expriment moins la subjectivité dans leur discours.

De plus, l'analyse des propriétés lexico-sémantiques et phraséologiques des éléments du LST en chinois permettra l'enrichissement de la ressource du LST en intégrant ces propriétés spécifiques au chinois afin de mieux guider le scripteur quant à leur emploi.

Enfin, les convergences et divergences repérées dans l'emploi du LST en français et chinois nous paraissent également intéressantes à confronter à un autre corpus d'écrits universitaires en français par des apprenants sinophones sur lequel nous avons déjà travaillé (Yan et Hatier, 2016).

Références bibliographiques

Cavalla, C. (2019). Une méthodologie sur corpus pour l'écriture en FOU. Points Communs - Recherche en didactique des langues sur objectif(s) spécifique(s), 47, 91.

Cavalla, C. (2021). Enseigner le lexique en classe de langue. Le Français dans Le Monde, CLE International, 435, pp.56-57. ⟨hal-03427888⟩

Chen, H. H. J., & Tao H. (2019). Academic Chinese: From Corpora to Language Teaching. In Xiaofei Lu and Berlin Chen (Eds.), Computational and Corpus Linguistic Approaches to Chinese Language Teaching and Learning (pp.57-79). Berlin & Singapore: Springer.

Coxhead, A. (2002). The academic word list: A corpus-based word list for academic purposes. Language and Computers, 42(1), 73–89.

Drouin, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. Revue française de linguistique appliquée, Vol. XII(2), 45 64.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP : Countering criticisms against corpus-based methodologies. English for Specific Purposes, 24(3), 321-332. https://doi.org/10.1016/j.esp.2004.09.002

Hatier, S., Augustyn, M., Yan, R., Tran, T. T. H., Tutin, A., & Jacques, M.-P. (2016). French cross-disciplinary scientific lexicon : Extraction and linguistic analysis. In T. Margalitadze & G. Meladze (Éds.), Proceedings of the XVII EURALEX International congress Lexicography & Linguistic diversity (p. 355-365). Ivane Javakhishvili Tbilisi State University.

Hu, G., & Cao, F. (2011). Hedging and boosting in abstracts of applied linguistics articles: a comparative study of English- and Chinese-medium journals. Journal of Pragmatics, 43(11), 2795–2809.

Kraif, O. (2016). Le lexicoscope : Un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. Cahiers de lexicologie: Revue internationale de lexicologie et lexicographie, 108, 91-106.

Li, Y. (李裕德) (1985). Grammar of scientific Chinese (科技漢語語法). Beijing: Metallurgical Industry Press (冶金工業出版社)

Loi, C. K., & Lim, J. M.-H. (2013). Metadiscourse in English and Chinese research article introductions. Discourse Studies, 15, 129-146.

Liu, C.(劉貞妤), Chen, H.(陳浩然), & Yang, H. (楊惠媚) (2016). Compiling a Chinese academic wordlist based on an academic corpus (藉學術語料庫提出中文學術常用詞表: 以人文社會科學為例). Journal of Chinese Language Teaching (華語文教學研究), 13(2), 4–87.

Mu, C., Zhang, L.-J., Ehrich, J. & Hong, H.-Q. (2015). The use of metadiscourse for knowledge construction in Chinese and English research articles. *Journal of English for Academic Purposes*, 20, 135-148.

Paquot, M. (2010). Academic vocabulary in learner writing: From extraction to analysis. Bloomsbury Publishing.

Paquot, M., & Bestgen, Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. Language and Computers, 68(1), 247 269.

Pecman, M. (2004). Phraséologie contrastive anglais-français: analyse et traitement en vue de l'aide à la rédaction scientifique (Thèse doctorat). Université de Nice-Sophia Antipolis. UFR Lettres, arts et sciences humaines, France.

Tao, H. (2013). Corpus of Written Academic Chinese. ACTFL CALPER Brochure. State College, PA: Pennsylvania State University.

Tran, T. T. H. (2014, décembre 11). Description de la phraséologie transdisciplinaire scientifique et réflexions didactiques pour l'enseignement à des étudiants non-natifs. Application aux marqueurs discursifs (Thèse de doctorat). Université de Grenoble, Grenoble.

Tran, T. T. H., & Royer, S. (2020). Analyse lexicométrique des collocations en hygiène et propreté et applications pédagogiques dans les formations pour les travailleurs migrants. Action Didactique. https://hal.science/hal-03197122

Tutin, A., & Kraif, O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents. Lidil. Revue de linguistique et de didactique des langues, 53, 119-141.

Wu, Ge Qi et Zhu, Yong Sheng (2014). Self-mention and authorial identity construction in English and Chinese research articles: A contrastive study. Linguistics and the human sciences, 10(2), pp. 133-158.

Yan, R., & Hatier, S. (2016). L'extraction et la modélisation de patrons lexico-syntaxiques pour leur enseignement en FLE : un exemple à partir du verbe montrer. Linguistik Online, 78(4), 93-112.

薛蕾.(2017).基于汉语语言学论文语料库的学术汉语词汇析取及特征研究(硕士学位论文, 云南师范大学).
https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201801&filename=1017730145.nh

黄启庆&薛蕾.(2018).汉语国际教育视角下的学术汉语词汇特点研究.世界汉语教学学会秘书处.(eds.)第十三届国际汉语教学研讨会论文选, 20-29. 商务印书馆 (The Commercial Press).

Constitution semi-automatique de corpus pour l'extraction et l'analyse des constructions causatives néologiques en -iser et en - 化[huà] dans le discours médiatique contemporain[127]

Jiahui Zhu 1 et David Kletz 2
1 Université Sorbonne Nouvelle & Lattice (CNRS/ENS-PSL/USN)
2 Université Paris Cité & LLF (CNRS/UPC)
jiahui.zhu@sorbonne-nouvelle.fr, david.kletz@sorbonne-nouvelle.fr

Introduction

Dans le cadre de notre étude visant à étudier les caractéristiques de la causalité liée au néologisme formé en *-iser* et en - 化*[huà]*, nous avons été amenés à construire un corpus de données écrites prove- nant des médias. En d'autres termes, notre étude est axée sur les néologismes suffixés en morphèmes causatifs *-iser* et - 化*[huà]* de deux langues génétiquement et typologiquement éloignées. La période ciblée est de 2000 à 2022. En nous appuyant sur l'étude contrastive des propriétés liées à la causalité exprimée par ces deux suffixes, nous essayons de répondre à la question suivante : comment les deux suffixes causatifs expriment la causalité dans ces vingt dernières années ? Comme le mentionne Habert *et al*. « [le corpus est utilisé pour] tester des hypothèses, confronter un modèle postulé aux réalisa- tions effectives» (1997 : 08). La constitution d'un corpus représente la phase préalable à l'analyse des données et joue un rôle incontournable, particulièrement pour une analyse linguistique qui porte sur deux langues aussi distinctes que le français et le chinois. Ainsi, *comment constituer un corpus pour les études contrastives des constructions néologiques causatives suffixées en -iser et en - 化[huà] de deux langues éloignées pour la période de 2000 à 2022* ? Pour répondre à cette question, cette présen- tation se concentrera sur les aspects méthodologiques de la constitution d'un corpus comportant des néologismes causatifs en *-iser* et en - 化*[huà]* apparus après l'an 2000 dans le discours médiatique et les concordances correspondantes, et sur l'état des données de recueillies.

Corpus et méthodologie

Pour constituer notre corpus de néologismes, nous avons fait le choix de réunir des données écrites dans le discours médiatique. Ce choix a été fait pour des raisons de taille, d'opérabilité de l'extraction des données et de représentativité des méta-données (Carter-Thomas, 2009, 112-117).

[127]- Cet article est financé par China Scholarship Council (CSC)

Plus précisément, les méta-données de cette recherche proviennent d'une part des médias traditionnels : les journaux français qui se trouvent dans la plateforme *Europresse*[128], *Quotidien du Peuple* (la presse nationale chinoise) [129] et *Quotidien du Sud* (la presse régionale chinoise) [130] ; d'autre part des médias d'aujourd'hui : *Twitter* [131] (cf. Daoust, 2017) pour le français et *Sina Weibo* [132] pour le chinois.

Avant de présenter les méthodes permettant de constituer un corpus concernant des néologismes causatifs dérivés du suffixe *-iser* et du suffixe - 化*[huà]*, il est nécessaire de commencer par la définition de ce type de constructions acceptée par notre recherche. Guidés par la Grammaire Cognitive de Construction (cf. Goldberg, 1995 ; 2006 ; Bouveret & Legallois, 2012 ; Carlier & Prévost, 2021) et la Morphologie de Construction (cf. : Booij, 2010), nous définissons le néologisme causatif suffixé par *-iser* et par - 化*[huà]* en tant que nouvelle construction verbale causative en *-iser* et en - 化*[huà]* apparue après l'an 2000. Cette construction est soit un lexème, soit un syntagme syntaxique / idiomatique et au moins trois occurrences sont avérées dans des ressources médiatiques ou encyclopédiques.

Nous appuyant sur cette définition, nous commençons la présentation des méthodes de constitution de ce corpus pertinent pour notre recherche.

Afin de constituer un corpus qui contienne des données aussi complètes que possible, dans un premier temps, nous avons besoin de sélectionner toutes les constructions comportant les morphèmes cibles en chinois et en français. Plus précisément, pour le chinois, nous avons sélectionné tous les termes finis par le caractère - 化*[huà]* ; pour le français, nous avons extrait d'une manière globale les termes qui se terminent par le morphème *-iser* et ses flexions. Cette étape est réalisée avec *Python*. Grâce aux librairies *Jieba* (Sun, 2012) et *Spacy* (Honnibal & Montani 2017) nous avons extrait respectivement 1427 candidats-termes chinois et 2672 candidats-termes français.

Dans un second temps, nous cherchons à obtenir deux listes ne contenant que les constructions causatives néologiques. En d'autres termes, nous avons besoin de supprimer les candidats-termes ne respectant pas les termes de notre recherche. Ce tri est effectué de manière semi-automatique

Dans notre étude, nous avons adapté la méthodologie de pré-traitement des données de Cartier & Huyghe (2021 : 6). Pour constituer un corpus des préfixes de *haut degré* en français, les auteurs ont éliminé les formes fautives et non préfixées, ainsi que les formes préfixées non pertinentes (*ibid*). Nous adaptons cette méthode de pré-traitement pour l'étude de la suffixation en éliminant semi- automatiquement :

[128]- https ://nouveau.europresse.com/

[129]- http ://www.people.com.cn

[130]- https ://epaper.southcn.com/nfdaily/html/202206/12/nodeA01.html

[131]- https ://twitter.com/home

[132]- https ://m.weibo.cn

- les formes fautives : il s'agit essentiellement des résultats de fautes d'orthographes (i.e. **démocraiser* en français). Ceci peut être éliminé automatiquement en s'assurant que le radical n'appartient pas au lexique de la langue cible.
- les formes non suffixées : celles qui comportent les chaînes de caractères recherchées, mais ne sont pas analysables morphologiquement (i.e. *viser* en français, 文化 [wén huà] (n.'culture') en chinois).
- les formes suffixées non pertinentes : les formes suffixées non pertinentes sont celles dans lesquelles les suffixes ne peuvent pas former la causalité. Ce type de cas se trouve essentiellement en chinois. Par exemple, à la lumière des études sur les méta-données, nous avons constaté que le terme suffixé 二胎化[èr tāi huà] (二胎化[èr tāi huà]< 二胎[èr tāi](n. 'deuxième enfant') + - 化suff.[huà]) ne peut être employé que comme substantif ou adjectif. [133]. En conséquence, nous avons besoin d'en référer aux concordances pour déterminer manuellement la catégorie grammaticale de la suffixation par - 化 [huà].
- les problèmes de tokenisation en chinois (i.e. 取冰化水[qǔ bīng huà shui] tokenisé en * 取冰化[qǔ bīng huà] + 水 [shuǐ] plutôt qu'en 取冰[qǔ bīng] + 化水[huà shuǐ] en chinois).

D'autre part, du point de vue de la néologie, nous avons besoin également de supprimer semi-automatiquement les hapax et les constructions apparues avant l'an 2000. Les hapax sont les termes qui se trouvent seulement une ou deux fois dans les ressources médiatiques et encyclopédiques (i.e. *bling-bling-iser*, qui se trouve uniquement une fois en français ; 精勤化 [jīng qín huà], qui n'est apparu qu'une seule fois en chinois). En ce qui concerne l'élimination des candidats-termes apparus avant l'an 2000, il est indispensable de dater leur première apparition. Puisqu'il n'existe pas de corpus établi et complet permettant de dater précisément l'apparition des termes chinois et français, dans cette étude, nous avons construit notre propre méthode pour essayer d'identifier les premières occurrences des constructions chinoises et françaises.

Pour dater la première apparition des constructions françaises, en nous appuyant sur des corpus existants, une base de données de référence a été créée, couvrant la période de 1500 à nos jours.

D'une part, nous nous sommes référés automatiquement à la plateforme *Google Ngram* (ci-après abrévié en *GNG*) (Michel *et al*., 2010 ; Lin *et al*., 2012). *GNG* a extrait des documents numérisés à partir de corpus *Google Books* en français contemporain au cours de la période 1500-2019. Les documents en question sont principalement des ouvrages littéraires et des documents scientifiques et techniques. Ce corpus fournit la fréquence d'utilisation des constructions au fil du temps et génère un graphique basé sur l'évolution de la fréquence. Ainsi, lorsque la fréquence d'un terme cesse d'être zéro à un certain moment du graphique de *GNG*, nous pouvons déterminer qu'il a commencé à apparaître.

D'autre part, nous nous sommes référés manuellement à la plateforme de journaux :*Europresse* et au média d'aujourd'hui :*Twitter*. *Europresse* englobe tous les principaux journaux français en Hexagone depuis 1840. *Twitter* nous aide à saisir la fréquence de l'utilisation des néologismes en français depuis 2006. Ces deux références sont aussi importantes pour décider les dates d'apparition des constructions pour deux raisons principales :

[133] Quand le dérivé chinois 二胎化 [èr tāi huà] s'emploie en tant que substantif, il signifie « un changement social chinois concernant une famille devenant deux enfants» ; lorsqu'il joue un rôle d'adjectif, il veut dire « qui est lié à ce changement social»

- Premièrement, lorsque certaines constructions sont assez récentes, *GNG* ne peut pas rechercher leur fréquence. Par exemple, nous ne trouvons pas de données sur le terme *chiciser* (ou *chic-iser*, ou *chiquiser*) dans le *GNG*.
- Deuxièmement, les genres discursifs représentés dans le *GNG* sont homogènes, provenant seule- ment de la littérature ou d'ouvrages scientifiques. Une datation précise de la première occurrence d'une construction française doit tenir compte de la représentativité des données de référence, c'est-à-dire avoir une vision globale et précise des premiers emplois des candidats-termes dans la société française. Pour atteindre cette représentativité, la référence à un seul genre discursif ne suffit pas : il faut une diversité des sources de données (Thuilier, 2012 : 45). Par conséquent, nous avons ajouté le genre du discours médiatique parmi les données de référence puisque les médias permettent une plus grande diversité et dans l'expression et dans la diffusion des infor- mations. En effet, les micro-blogs qui sont écrits par des auteurs de régions différentes, exerçant des professions différentes, représentant des attitudes différentes, s'exprimant de façon distincte, sont à même de répondre à la diversité des sources requise par notre collecte de données. Pa- rallèlement, le caractère instantané des médias nous permet de déterminer de façon précise la première apparition des néologismes. Les médias, surtout les nouveaux médias, constituent des outils puissants capables de capturer à temps réel les néologismes qui émergent à mesure que la société évolue. La capacité des réseaux sociaux d'enregistrer l'évolution de la société offre la possibilité de relever les premières occurrences d'un néologisme. En définitive, la pertinence du genre discursif des médias, en particulier des réseaux sociaux, émane de sa diversité des sources, de son caractère instantané et de la facilité de dater avec précision la première occurrence des candidats-termes cibles.

Après avoir effectué ces trois extractions, nous avons obtenu trois résultats complémentaires sur la datation, que nous avons introduits dans le tableau *csv*. La plus ancienne de ces trois données a été déterminée comme étant la première occurrence de la construction.

Pour la datation des premières apparitions des constructions chinoises, en nous appuyant sur des journaux et de réseau social chinois, nous avons organisé une base de données de référence couvrant la période de 1840 à nos jours.

D'une part, nous avons choisi *Sina Weibo*. Néanmoins, étant donné que ce réseau social n'est publié que depuis 2009, ce n'est qu'après cette année que nous sommes en mesure d'y observer la fréquence de l'utilisation des constructions chinoises.

D'autre part, nous avons utilisé les plateformes de presses comme références. La première plate- forme de journaux que nous avons choisie est *Journal Moderne en Chine* (近代报纸数据库 [jìn dài bào zhǐ shù jù kù]) [134], qui offre un accès électronique à 208 journaux couvrant la période 1840-1949. La deuxième plateforme sur laquelle nous nous sommes appuyés est *Librairie Numérique Chinoise* (中华数字书苑 [zhōng huá shù zì shū yuàn]) [135], nous permettant d'accéder à 456 grands journaux électro- niques en chinois publiés depuis 1946. La troisième plateforme à laquelle nous nous sommes reportés est *Base de données en texte intégral de journaux principaux chinois* (中国重要报纸全文数据库[zhōng guó zhòng yào

[134]- http ://tk.cepiec.com.cn

[135]- http ://www.apabi.com/jigou ?pid=about&cult=CN

bào zhǐ quán wén shù jù kù]) [136], regroupant des articles de 618 journaux chinois, datant de l'an 2000 à nos jours. Après avoir exécuté les extractions distinctes dans le réseau social et dans les journaux, nous avons disposé de deux résultats complémentaires sur la date d'apparition, que nous avons introduits dans le tableau *csv*. La plus ancienne de ces deux données a été identifiée comme étant la date de la première apparition de la construction correspondante.

La segmentation des articles réalisée par *Spacy* et *Jieba* afin d'extraire automatiquement toutes les constructions néologiques à étudier nécessite de s'appuyer sur une liste de références. Ainsi, après avoir reçu les listes des constructions néologiques causatives suffixées, nous l'avons convertie au format *txt.* pour utiliser ce fichier comme référence.

Finalement, nous avons importé les données segmentées dans le concordancier TXM (Heiden *et al.*, 2010), récupéré et exporté les concordances liées à chaque construction dans un tableau *csv.* pour les analyses linguistiques.

Résultats

Les contraintes induites par l'utilisation de données secondaires ou par le manque de bases données publiées ont pu être évitées grâce à la constitution de notre propre corpus. Nous avons sélectionné automatiquement toutes les constructions néologiques possibles suffixées en *-iser* et en - 化 *[huà]* apparues après l'an 2000 dans le discours médiatique. Puis, nous avons supprimé semi-automatiquement les constructions qui ne sont pas pertinentes pour notre objet de recherche. Notre liste totalise finalement 1200 constructions verbales néologiques françaises et 700 chinoises. En outre, nous avons extrait grâce à TXM 24 000 concordances françaises et 9278 concordances chinoises. Ces données nous permettront d'étudier la causalité exprimée par les constructions néologiques suffixées en *-iser* et en - 化 *[huà]*

Notre méthode se veut systématique et précise et les résultats obtenus révèlent sa pertinence et sa précision. Elle pourrait être utilisée dans le cadre d'autres recherches puisqu'elle propose un nouveau procédé productif pour sélectionner et identifier les néologismes qui apparaissent dans une période spécifique.

En outre, elle permettrait une étude en micro-diachronie concernant des suffixes *-iser* et - 化 *[huà]*.

L'approche de la micro-diachronie a été proposée en opposition à la dichotomie diachronie/synchronie suggérée par Saussure. Dans l'étude synchronique, Saussure considère le langage comme un système fermé et immobile, mais rejette l'idée d'une évolution possible dans une période donnée courte (cf. Saussure 1995) ; l'étude diachronique se concentre sur les processus de changement qui se sont produits dans l'histoire évolutive de la langue (ibid). L'étude traditionnelle en diachronie prend en compte des périodes très longues, des phénomènes linguistiques qui impliquent des décennies, voire des siècles (Dury, 2021 : 02). Toutefois, il est difficile de démentir que les langues se modifient à tout moment, ce que mentionne Cartier : « les langues sont des systèmes qui à la fois sont extrêmement stables dans le temps, mais qui sont aussi en continuelle évolution » (2018 : 186). Bien que la langue doive garder une stabilité pour assurer la communication, elle se modifie également sans cesse

[136]- https ://kns.cnki.net/kns/brief/result.aspx ?dbprefix=CCND

en réponse aux changements de la société (Feuillard, 2001 : 07). Il est ainsi raisonnable de proposer que le langage soit également susceptible de se modifier dans un court laps de temps. Contrairement à l'affirmation de Saussure selon laquelle une étude à court terme ignore les changements linguistiques, la micro- diachronie s'intéresse davantage aux changements en mouvance à l'intérieur du système linguistique dans un délai court déterminé (cf. Siouffi *et al*. 2012). Tels que les changements linguistiques résultant du développement social ou du contact linguistique au cours de ce siècle. En fonction de la qualité des données disponibles, l'empan chronologique de l'étude en micro-diachronie peut varier de dix ans à cinquante ans (Abouda & Skrovec, 2022 : 10). C'est en nous appuyant sur cette approche, la micro-diachronie, que nous pouvons étudier le changement linguistique lié à la construction causative suffixée pendant les deux dernières décennies. La méthode de constitution de corpus présentée offre ainsi la possibilité d'une étude en dynamique linguistique dans le discours médiatique (cf. Drescher, 2015) durant ces vingt dernières années. Un corpus annoté avec la première apparition des constructions causatives nous permet d'une part d'étudier une variation dans le choix des racines avec lesquelles ces suffixes sont combinés et d'autre part d'observer un changement dans la valeur causative depuis le début de ce siècle.

Références bibliographiques

Abouda, L. & Skrovec, M. (2022). « Micro-diachronie de l'oral. Présentation », Langages, vol. 226, no. 2, 2022, p. 9-24. Booij, G. (2010). Construction Morphology. New York : Oxford University Press.

Bouveret, M. & Legallois, D. (2012). « Cognitive linguistics and the notion in French studies : An overview », dans *Constructions in French*, Amsterdam / Philadelphia, John Benjamins Publishing Company, p. 1-22.

Carlier, A. & Prévost, S. (2021). « Constructions, constructionalisation et changement linguistique, présentation », dans Langue Française, n°209, p. 9-22.

Carter-Thomas, S. (2009). Texte et contexte : pour une approche fonctionnelle et empirique, HDR, Université de la Sorbonne Nouvelle.

Cartier, E. (2018). « Dynamisme lexical des langues : éléments théoriques, méthodes automatiques, expérimentations en français contemporain », Habilitation à Diriger des Recherches, Université Paris 13.

Cartier, E. & Huyghe, R. (2021). « La concurrence affixale en diachronie : le cas des préfixes de haut degré en français». Linx, p. 1-25.

Daoust, J.-F. (2017). «Démocratisation de l'information : effets différenciés des médias traditionnels et des nouveaux médias», dans Politique et sociétés, Volume 36, N°1, p.25-46.

Drescher, M. (dir.) (2015). Médias et dynamique du français en Afrique subsaharienne. Francfort-sur-le-Main : Peter Lang.

Dury, P.(2021). « L'obsolescence terminologique dans le domaine de la pharmacologie », Linx [En ligne], 82 | 2021, mis en ligne le 15 juillet 2021, consulté le 01 juin 2023.

Ferdinand de Saussure, Cours de linguistique générale(1916). Bailly, Ch. et Séchehaye, A. (éds.), Paris : Payot, 1995. Feuillard, C. (2001). « Le fonctionnalisme d'André Martinet », La linguistique, vol. 37, no. 1, 2001, p. 5-20.

Goldberg, A.-E. (1995). Constructions : A Construction Grammar Approach to Argument Structure. Chicago : Chicago University Press.

Goldberg, A.-E. (2006). Constructions at work : The nature of generalization in language Oxford : Oxford University Press.

Habert B., A. Nazarenko & A. Salem. (1997). Les linguistiques de corpus. Armand Colin : Paris.

Heiden, S., Magué, J.-P. & Pincemin, B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement», dans Sergio Bolasco, Isabella Chiari, Luca Giuliano (Eds.), Proc. of 10th In- ternational Conference on the Statistical Analysis of Textual Data - JADT 2010 (Vil. 2, p. 1021-1032). Edizioni Universitaire di Lettere Economia Diritto, Roma, Italy.

Honnibal, M., & Montani, I. (2017). spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Lin, Y., Michel, J.-B., Lieberman Aiden , E., Orwant, J., Brockman, W. & Petrov, S.(2012). «Syntactic Annotations for the Google Books Ngram Corpus», in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics Volume 2 : Demo Papers (ACL 2012)

Siouffi G., Steuckardt A. & Wionet C. (2012). «Comment enquêter sur des diachronies courtes et contemporaines ?», dans F. Neveu et alii (éds), 3e Congrès Mondial de Linguistique Française –CMLF 2012 (Lyon, France), SHS Web of Conferences 1, Les Ulis, EDP Sciences, p. 215-226

Michel, J.-B., Kui Shen, Y., Presser Aiden, A., Veres, A., Gray, M., Brockman, W. & the Google Books Team. (2010). « Quantitative Analysis of Culture Using Millions of Digitized Books», Science 331 (6014), p. 176-182.

Sun, J.(2012). «Jieba chinese word segmentation tool.» Accessed : Jun 25, p. 2018.

Thuilier, J. (2012). « Contraintes préférentielles et ordre des mots en français » (Thèse de doctorat de la science du langage), Université Paris-Diderot.

Base de données en texte intégral de journaux principaux chinois (China Core Newspapers Full-text Database) : https ://chn.oversea.cnki.net/kns ?dbcode=CCND

Europresse : https ://nouveau-europresse-com

Journal Moderne en Chine (Moderne Newspaper in China) :

http ://2.829.a697.unv.proxy1.online/library/publish/default/Main.jsp

Librairie Numérique Chinoise (Chinese Digital Library) : http ://www.apabi.com/jigou/pub.mvc/Index2 ?pid=login & cult=CN

Qutodien du Peuple : http ://www.people.com.cn Quotidien du Sud :https ://epaper.southcn.com Sina Weibo : https ://m.weibo.cn

Twitter : https ://twitter.com

Annexe

Nous avons dressé le tableau I afin de montrer les catégories ainsi que les tailles des sources des méta-données :

TABLE 1 – Origine des méta-données des corpus

Titre	**Catégorie**	**Taille**
Europresse	Plateforme comportant les presses nationales et régionales principales en Hexagone	26,405,733 tokens
人民日报 ([rén mín rì bào] *Quotidien du Peuple*)	Presse nationale chinoise	6,408,448,592 tokens
南方日报 ([nán fāng rì bào] *Quotidien de Nanfang*)	Presse régionale chinoise	1,231,717,562 tokens
Twitter	Nouveaux médias en France	754, 468 tokens
Sina Weibo	Nouveaux médias en Chine	9, 153, 937 tokens

Vers l'intégration des outils d'annotation syntaxique : proposition d'une chaîne de traitement itérative pour faciliter l'adoption et l'accès aux technologies d'apprentissage automatique

Rayan Ziane[1], Natalia Romanova[1]
[1]Laboratoire CRISCO, Université Caen
rayan.ziane@unicaen.fr

Ces dernières années on assiste à l'émergence d'outils d'intelligence artificielle avec apprentissage automatique pour l'analyse syntaxique, sous l'impulsion de chercheurs qui les rendent librement disponibles à la communauté. Un exemple en est l'analyseur syntaxique HOPS (HOnest Parser of Sentences)[137] (Grobol et Crabbé, 2021). Le parseur fournit a) une annotation rapide et de haute qualité[138] de corpus en français ancien et moderne grâce à des modèles de langues pré-entraînés et b) un entraînement de modèles d'autres langues et/ou de types de textes particuliers à partir d'échantillons pré-annotés par l'utilisateur. En suivant le programme d'annotation (*tagging* en parties du discours et *parsing* syntaxique) proposé par la communauté Universal Dependencies (UD) (de Marneffe et al., 2021) qui a déjà constitué des *treebanks* de plus de 100 langues différentes, dont de nombreuses langues peu dotées, l'analyseur permet aux utilisateurs d'inscrire leur travail dans le cadre d'une initiative internationale en plein développement qui vise à produire un système stable et cohérent pour décrire la variation inter- et intralinguistique. En outre, l'annotation syntaxique en UD, tout en ouvrant de nombreuses possibilités de recherche et d'application du corpus résultant, peut servir de base pour d'autres couches d'annotation.

Le but de la présentation est d'articuler des réponses aux défis d'une analyse intégrée des données textuelles. Cette proposition est faite au regard des obstacles qui subsistent pour l'adoption généralisée des outils comme l'analyseur HOPS, notamment en ce qui concerne les chercheurs travaillant au croisement de la linguistique, la constitution de corpus en diachronie et l'édition numérique. Même si, après l'installation, HOPS est relativement facile à prendre en main, il existe nombre de défis pour l'intégration de cet outil dans le processus du traitement des données. Premièrement, l'utilisation de l'analyseur présuppose un prétraitement qui implique un programme de segmentation du texte à l'entrée en unités maximale et minimale d'analyse (en phrases et en *tokens*). La segmentation et la longueur de la phrase ont un impact direct sur la qualité de l'annotation produite en sortie (Grobol *et al.*, 2021). Cependant, la réalité matérielle des données souvent hétérogènes, provenant de différentes sources, comme par exemple des versions numérisées des imprimés du 16ème siècle

[137] https://github.com/hopsparser/hopsparser

[138] Le modèle SRCMF-UD pour l'ancien français ayant un taux de précision entre 96 et 98% sur les textes du corpus sur lequel il a été entraîné (https://github.com/hopsparser/hopsparser/blob/main/docs/models.md).

ou des transcriptions semi-diplomatiques des sources manuscrites anciennes et modernes, rend la segmentation difficile à contrôler et à uniformiser (Goux & Pinzin, sous presse). Les pratiques des éditeurs et des imprimeurs divergeant considérablement, les ponctuations ni la découpe en mots typographiques ne sont pas toujours fiables et cohérentes à l'intérieur du corpus, voire au sein de chaque texte individuellement. Deuxièmement, le parseur HOPS prévoit l'utilisation du format CONLL-U, adopté par UD, tandis que le format XML-TEI (de Rose, 1999) reste le format particulièrement adapté aux pratiques de l'édition numérique. En outre, les projets ont des buts de recherche et de dissémination différents, auxquels la contrainte de l'annotation UD peut ne pas correspondre. Finalement, malgré le taux plus qu'impressionnant de succès de l'annotation, l'utilisation de HOPS sur des textes non-littéraires en ancien et moyen français nécessite toujours un programme de vérification manuelle.

A l'heure actuelle, en vue des progrès de l'application des outils d'apprentissage automatique dans le domaine de la recherche universitaire en linguistique, notamment en syntaxe, et le rôle croissant des communautés internationales autour des initiatives comme Universal Dependencies qui cherchent à promouvoir une approche standardisée à l'analyse des langues sans évincer les inévitables dimensions de variation, nous constatons un réel besoin de définir des protocoles qui faciliteraient l'adoption des nouveaux outils et approches par le plus grand nombre de chercheurs dans un contexte le plus interdisciplinaire possible. L'accès aux nouvelles technologies, à son tour, garantira la longévité des initiatives et outils et leur développement grâce aux retours des collègues et aux données annotées générées par les nouveaux projets qui pourront être réutilisées pour améliorer les outils.

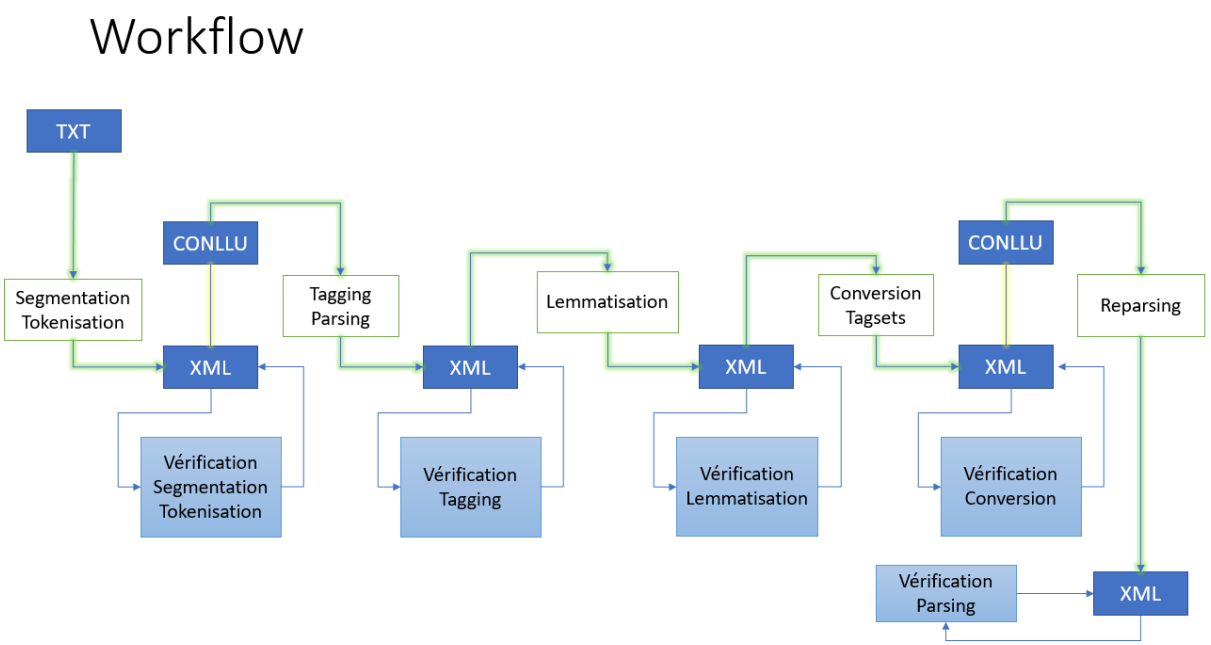
La communication proposée portera donc sur une chaîne d'annotation et lemmatisation semi-automatique intégrant l'utilisation de l'analyseur syntaxique HOPS.[139] La chaîne de traitement High-Tech-CRISCO, dont une version bêta des scripts Python et la documentation ont été mis à disposition via un projet GitHub en juin 2023,[140] est développée et testée dans le cadre du projet High-Tech (High-Level Text Annotation across Historical Texts : Improving semi-automatisation of big-data management)[141] et est testée lors de la création de la partie française du corpus du projet MICLE (MICro-Indicateurs de l'Évolution grammaticale : un modèle multifactoriel de la perte de V2 en vénitien et en français anciens) au laboratoire CRISCO (Université de Caen). Ces deux projets, dont les équipes réunissent des compétences en traitement automatique des langues (TAL), en syntaxe et diachronie du français, en édition numérique et en paléographie, portent sur la création de deux corpus en diachronie, calibrés par provenance géographique et genres (non-littéraires) des textes traités. Les deux corpus réuniront une vingtaine d'échantillons en français d'autour de 40,000 *tokens* chacun, allant du 13ème au 19ème siècles, annotés syntaxiquement et lemmatisés.[142] En plus de l'annotation PoS

[139] Cette chaîne pourrait évidemment être adaptée pour l'utilisation d'un autre analyseur syntaxique disponible, par exemple Stanford *parser* (https://nlp.stanford.edu/software/lex-parser.shtml).
[140] https://github.com/RZiane/HT_CRISCO.

[141] Le projet ANR-DFG MICLE est financé pour la période 06/2021-05/2024. (https://www.unicaen.fr/projet_de_recherche/micle/) et le projet RIN (Réseau d'Intérêts Normands) High-Tech pour 12/2021-2023. Vu la disponibilité des modèles pré-entraînés pour le parseur HOPS, seul le corpus français du projet MICLE est annoté en utilisant la chaîne de traitement présentée ici.

[142] Une première version du corpus français MICLE et du corpus High-Tech est consultable via le portail TXM du CRISCO https://txm-crisco.huma-num.fr/txm/ depuis avril 2023.

en UD qui offre des catégories assez générales, la chaîne prévoit une conversion en deux autres jeux d'étiquettes, UPenn (Santorini, 2007) et Presto (Blumenthal *et al*, 2017), ce qui non seulement permettra d'affiner le grain de l'analyse mais aussi favorisera l'interopérabilité avec d'autres corpus et le dialogue avec des chercheurs à l'échelle nationale et internationale.



- **La chaîne d'annotation syntaxique**

La chaîne de traitement consiste en cinq phases : 1) le prétraitement (segmentation et tokenisation) 2) *tagging* et *parsing* UD (HOPS) 3) lemmatisation (dictionnaire Presto) 4) conversion des jeux d'étiquettes (UPenn et Presto) 5) reparsing du texte annoté et vérifié (HOPS). La phase finale prend en compte les modifications apportées à la tokenisation et à la segmentation en phrase au cours des phases 3-4 ; les étiquettes UD obtenues à cette étape ne sont normalement pas retenues sauf pour une éventuelle évaluation.

Pour toutes les phases, la chaîne de traitement propose des simples interfaces faciles à utiliser pour effectuer la sélection des fichiers d'entrée et d'autres options. Par exemple, lors de l'exécution de Phase 1 (pré-traitement), l'utilisateur a une série d'options pour la segmentation en phrases d'après des ponctuation fortes et peut obtenir des statistiques sur la longueur des phrases à la sortie et une liste de phrases jugées trop longues qui doivent être segmentées manuellement ().

Le principe de la chaîne de traitement est de mettre à profit les éléments déjà acquis à une étape précédente pour y ajouter de nouvelles métadonnées : par exemple, la conversion des jeux d'étiquettes repose sur l'annotation PoS UD, d'un côté, et, de l'autre, sur la lemmatisation qui, elle-même, dépend de l'analyse UD. Une étape de vérification manuelle est prévue entre chacune des cinq phases, ce qui permet de contrôler les erreurs et ne pas les laisser se propager sur les phases suivantes. Des corrections peuvent donc être apportées à la transcription, tokenisation, segmentation en phrases, lemmatisation et annotation en PoS à

n'importe quelle étape du traitement. Cette approche favorise la flexibilité, enlève la pression sur la phase de prétraitement et donne aux chercheurs la liberté de pouvoir modifier leur analyse tout en se familiarisant avec le texte traité (Morcos et al, 2021).

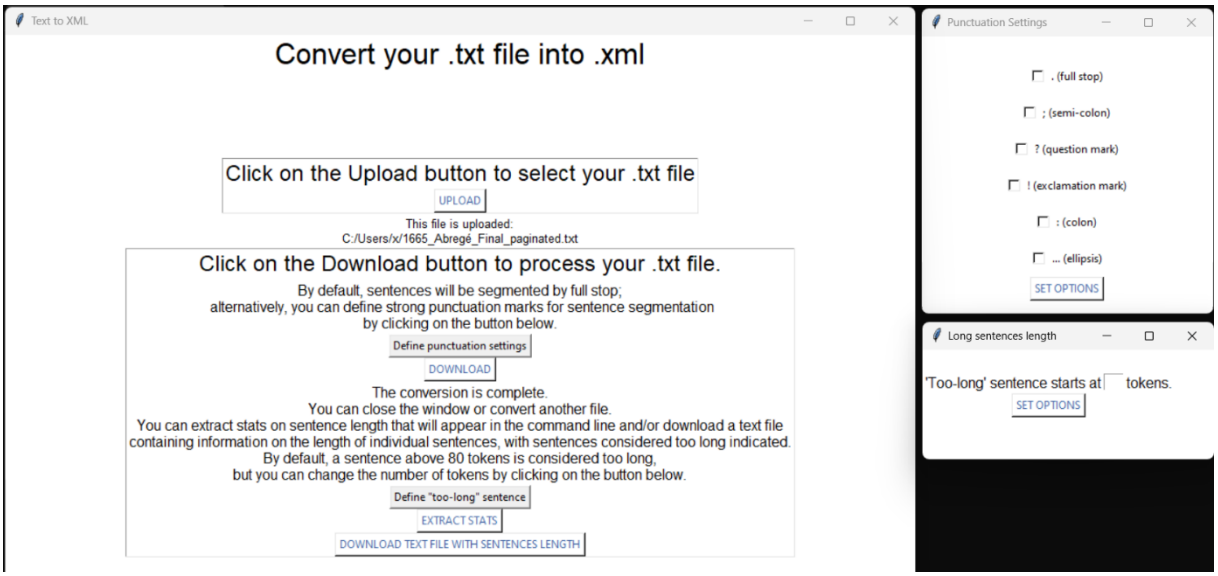


figure . 3 **Phase 1, conversion XML/TEI, segmentation en phrases et tokenisation (interface)**

La chaîne de traitement repose sur l'utilisation du format XML-TEI et fait appel aux outils et ressources déjà disponibles (dont l'analyseur syntaxique HOPS) et une série de scripts développés dans le cadre du projet High-Tech. Faciles à exécuter à partir de la ligne de commande ou via un environnement comme GitHub Desktop, les scripts sont utilisés pour assurer les différentes phases la chaîne ; nous proposons également une série de scripts outils, intégrés dans les scripts des phases principales mais qui peuvent être utilisés séparément (par exemple pour la conversion en format CONLL-U, pour la numérotation des *tokens* et des phrases et la désambiguïsation des parties du discours après conversion en UPenn et Presto). Pour faciliter la vérification, les scripts sont disponibles pour générer des tableurs pour la révision des formes, PoS et lemmes.

Lors de la communication, nous allons exposer les différentes étapes de la chaîne de traitement en utilisant les exemples tirés des deux textes anglo-normands du corpus MICLE datant de la fin du 13ème et de la première moitié du 14ème siècle. Ces textes présentent un défi particulier parce que l'anglo-normand, le français utilisé en Angleterre au Moyen Âge se distingue par une forte variation, notamment lexicographique. Pour pouvoir compléter la chaîne, nous avons dû recourir à une couche de lemmatisation supplémentaire en lemmes anglo-normands pour pouvoir ensuite les convertir en français dit standard. Cette duplication de l'annotation a l'avantage de montrer des potentielles applications de la chaîne en dehors de l'analyse syntaxique, par exemple en lexicographie, ce qui favorise une collaboration interdisciplinaire et réutilisation des données.

Au cours du développement de la chaîne High-Tech-CRISCO, nous avons eu de multiples opportunités d'échange avec des collègues du laboratoire et des autres universités en France et

à l'étranger qui souhaiteraient intégrer l'annotation syntaxique et la lemmatisation dans leurs projets et cherchent un protocole simple et facile à suivre qui serait adapté à leur pratique. Inspirée par ces discussions et retours, la chaîne de traitement proposée s'inscrit donc dans le courant de l'adoption des outils d'apprentissage automatique dans la recherche linguistique, édition numérique et constitution de corpus écrits. Elle favorisera la FAIRisation (Wilkinson et al, 2016) ainsi que l'ouverture des données de la recherche tout en facilitant l'accès à une annotation fiable et facile à prendre en main et encourageant l'interopérabilité et la réutilisation des corpus en dehors et post-projet.

Références bibliographiques

Blumenthal, P., Diwersy, S., Falaise, A., Lay, H., Souvay, G., Vigier, D., Descartes, P. R., Nancy, U., de Lyon, U. (2017). *Presto, un corpus diachronique pour le français des XVIe-XXe siècles*.

de Marneffe, M.-C., Manning, C. D., Nivre, J., Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255-308. https://doi.org/10.1162/coli_a_00402

DeRose, S. (1999). XML and the TEI. *Computers and the Humanities*, *33*(1), 11-30. https://doi.org/10.1023/A:1001771114509

Goux, M., Pinzin F. Challenges of a Multilingual Corpus (Old French/Old Venetian): The Example of the MICLE project. *Venise et la France. Similitudes, spécificités, interrelations*. Castro E., Della Fontana A. and Pezzini E. Franco Cesati (eds) Florence : Cesati Editore (sous presse).

Grobol, L., Crabbé, B. (2021). Analyse en dépendances du français avec des plongements contextualisés (French dependency parsing with contextualized embeddings). *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, 106-114. https://aclanthology.org/2021.jeptalnrecital-taln.9

Grobol, L., Prévost, S., Crabbé, B. (2021). Is Old French tougher to parse? *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, 27-34. https://aclanthology.org/2021.tlt-1.3

Lay, M.-H., Pincemin, B. (2010). Pour une exploration humaniste des textes : AnaLog. Statistical Analysis of Textual Data: *Proceedings of 10th International Conference Journée d'Analyse statistique des Données Textuelles* 9-11 June 2010 – Sapienza University of Rome. Bolasco, S., Chiari I. & Giuliano L. (eds) V.2, 1045-1056 https://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-1045-1056_106-Lay.pdf

Morcos, H., Noël, G., Husar, M. (2021). Lemmatization in the collaborative editorial workflow of a medieval French text: The digital edition of the *Histoire ancienne jusqu'à César*. *Digital Scholarship in the Humanities*, 36(2), 203-209. https://doi.org/10.1093/llc/fqaa060

Santorini, B. (2007). *Protocole d'étiquettage – Parties du discours (PDD)*. https://www.ling.upenn.edu/~beatrice/corpus-ling/annotation-french/pos/pos-index.html

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Art. 1. https://doi.org/10.1038/sdata.2016.18

Didactique de la méthodologie de corpus et applications pratiques dans le tourisme patrimonial et en lexicographie bilingue spécialisée : le projet UniVOCIttà

Valeria Zotti[1]
[1] Dipartimento LILEC, Università Alma Mater Studiorum di Bologna, Italie

valeria.zotti@unibo.it

Introduction

L'objectif de cette contribution est de souligner l'intérêt de la création et de l'exploration de corpus comparables spécialisés dans la formation de niveau master en lexicographie bilingue et traduction dans le domaine du tourisme patrimonial. Nous illustrerons l'expérience menée avec une classe de Linguistique Française du Master international en Language, Society & Communication, ayant pour objectif la création d'un corpus spécialisé sur le patrimoine artistique italien, dont des extraits seront consultables sur une nouvelle application web mobile conçue pour le tourisme culturel (Fig.1). Cette application contiendra aussi un dictionnaire bilingue spécialisé (français-italien) dans le domaine des Beaux-Arts réalisé par la classe à partir de l'exploration du dit corpus spécialisé ainsi que d'autres corpus comparables.

Corpus et méthodologie

Corpus

La classe de linguistique française a collaboré activement au déroulement du projet de recherche interuniversitaire *Lessico plurilingue dei Beni Culturali* (https://www.lessicobeniculturali.net/) dont l'objectif principal est de fournir aux professionnels de l'art des ressources fiables pour la description plurilingue du patrimoine artistique italien, de la Toscane en particulier et de la ville de Florence. Dans le cadre de ce projet, un corpus multilingue (italien, français, anglais, espagnol, russe, allemand) a été constitué et continue d'être enrichi (http://corpora.lessicobeniculturali.net/).

Les étudiant.e.s ont été chargés, dans une première phase, de collecter et numériser les textes des trois catégories textuelles sélectionnées par le projet LBC (Billero 2020 ; Farina, Nicolas Martinez 2020) afin de constituer un sous-corpus sur une autre région italienne, limitrophe de la Toscane : l'Emilie-Romagne et son chef-lieu Bologne. Ce sous-corpus Bologne et Émilie-Romagne (BER) Français est constitué à l'heure actuelle de plus de 200 textes en français langue originale (aucune traduction), contenant plus de 700 000 tokens. Ce corpus est donc destiné à accroitre le corpus LBC Français d'environ 20% et à compléter la description de la terminologie artistique donnée par le corpus monolingue LBC Français.

La constitution du sous-corpus BER est partie intégrante d'un nouveau projet satellite de LBC qui a obtenu en 2022 un financement destiné à répondre à l'objectif de 'troisième mission' de l'Université, communément appelé « service à la société » : le projet *UniVOCIttà: Voci digitali sull'unicità del patrimonio bolognese* (Voix numériques sur l'unicité du patrimoine de Bologne). Bien que les deux projets visent la valorisation du patrimoine local, UniVOCIttà se distingue sur le plan des applications pratiques et du produit envisagé qui s'adressera d'abord au grand public et à l'industrie du tourisme plutôt qu'à un public spécialisé de linguistes.

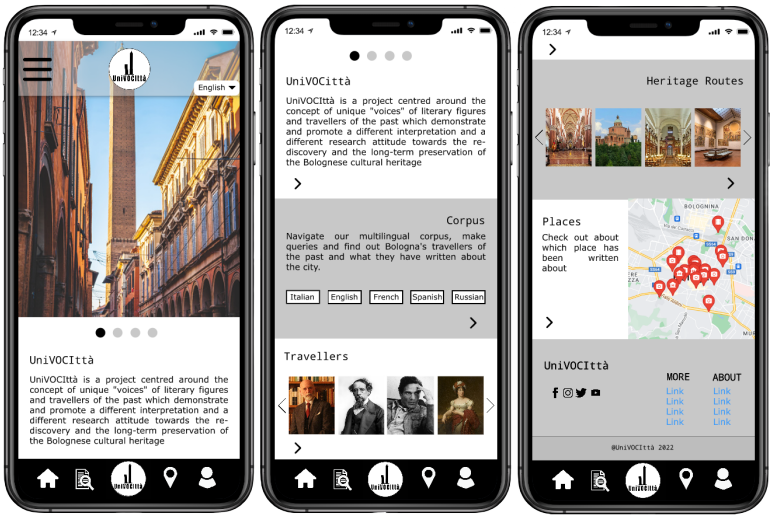


figure . 1 Prototype de l'application web mobile – corpus

Le produit final du projet sera un site web et une application web mobile (Fig. 1) qui permettra de consulter le corpus multilingue sous forme de fragments textuels décrivant le patrimoine de la ville de Bologne et de sa région à partir des témoignages laissés par d'illustres voyageurs étrangers du passé (hommes de lettres comme Montesquieu, Stendhal, Giono mais aussi scientifiques, astronomes, botanistes, etc.). Ces fragments textuels ont été étiquetés par mots-clés correspondant à des catégories thématiques par les étudiant.e.s avec le logiciel Atlas.ti, un logiciel performant pour l'analyse qualitative des données textuelles (Williams 2020 : 200).

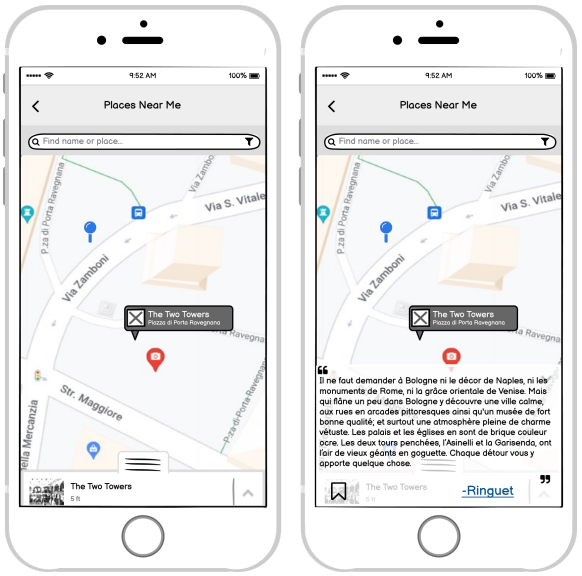


figure . 2 Prototype de l'application web mobile – géolocalisation avec extraits des corpus

Cette application contiendra aussi un dictionnaire bilingue spécialisé (français-italien) dans le domaine des Beaux-Arts réalisé par la classe à partir de l'exploration du corpus spécialisé BER ainsi que d'autres corpus comparables, ce qui correspond aux autres phases du travail proposé à la classe de linguistique française et traduction spécialisée.

Méthodologie

Dans les différentes étapes de l'expérience proposée en cours, les étudiant.e.s se sont occupé.e.s de:

Phase 1 : Collecte et étiquetage

1. La collecte des textes pour la constitution du corpus BER, avec des méthodes traditionnelles (conversion de textes repérés dans des bibliothèques numériques ou numérisation des versions papier) ou automatisées (BootCat) ;
2. L'étiquetage des fragments textuels qui figureront sur l'application mobile sur la base de critères qualitatifs (Atlas.ti) ;

Phase 2 : Extraction et triage

3. L'extraction terminologique (fonction « Keywords & Terms » de SketchEngine (https://www.sketchengine.eu) afin d'obtenir un échantillon de termes candidats représentatifs des discours sur la langue de l'art qui formeront la nomenclature du futur dictionnaire bilingue spécialisé ; dans cette phase, les problèmes courants des extracteurs de termes (L'Homme 2022) sont ressortis de manière évidente, à savoir les listes générées à la suite de l'extraction automatique ont renfermé des suites de mots qui n'intéressent pas l'utilisateur ;
4. L'ordonnancement des termes extraits : les candidats, triés par ordre alphabétique et par fréquence décroissante, sont ordonnés dans une liste est soumise à l'approbation des utilisateurs/apprenants.

Phase 3 : Validation et enrichissement

5. La vérification de la pertinence des candidats termes dans des dictionnaires de référence extensifs (TLFI, marques de domaine relatives aux Beaux-Arts) et dans des dictionnaires spécialisés (Félibien 1676, Viollet-Le Duc 1854, etc.) ;
6. L'exploration du corpus BER (fonction « Concordances ») afin de repérer des collocations spécialisées en utilisant aussi la fonction CQL pour la recherche des constructions complexes ;

Phase 4 : Analyse contrastive et traduction

7. L'analyse contrastive, c-à-d. examiner les termes et les collocations spécialisées qui posent des difficultés de traduction (ex. différents équivalents traductionnels selon le niveau de spécialisation du texte ; collocations spécialisées opaques), aussi à travers la consultation des principales ressources lexicographiques et terminologiques plurilingues et de deux traducteurs automatiques ;
8. La rédaction d'une entrée de dictionnaire bilingue en utilisant les corpus comparables pour la recherche des équivalents traductionnels, et, faute de données satisfaisantes, en validant les hypothèses traductives par la consultation de la Toile comme Corpus.

Résultats

Cette expérience didactique s'insère ainsi dans une réflexion plus vaste sur les méthodologies à adopter pour former des lexicographes bilingues et de futurs traducteurs, mais aussi sur le concept de compétence stratégique développé par Amparo Hurtado Albir (2008, 2019) et celui

de *Data-Driven Learning* - ou Apprentissage sur Corpus - proposé par Boulton et Tyne (2014). A travers l'application des principes pédagogiques du *Data Driven Learning*, nous montrerons que les principes actuellement défendus en didactique des langues (*Learning by Doing*) sont applicables à l'enseignement de la traduction spécialisée et que l'utilisation systématique de corpus textuels spécialisés, tels que les corpus comparable LBC et le corpus monolingue BER, permet d'améliorer la formation en traduction spécialisée et la qualité globale des traductions du lexique artistique dans des dictionnaires bilingues.

Un grand nombre d'études sur les corpus dans le domaine didactique portent sur des corpus de productions d'apprenants (« learner corpora », cf. Gandin 2016 ; Turci & Aragrande 2020 ; entre autres). L'approche que nous présentons ici se distingue de celle-ci, car les corpus en question ne contiennent pas les productions linguistiques des élèves eux-mêmes. Aussi, nous ne nous sommes pas limités à fournir aux étudiants des corpus déjà constitués, mais en leur demandant, dans la première phase, de constituer un corpus ciblé en utilisant le logiciel BootCat nous avons stimulé leur réflexion sur les critères à choisir pour la sélection des *seeds* (mots-clés). Aussi, face aux limites constatées de cette méthodologie pour le domaine envisagé, nous les avons impliqués dans la recherche de stratégies alternatives pour exploiter au mieux le logiciel de manière plus efficace.

Dans les phases suivantes, les étudiant.e.s ont expérimenté dans quelle mesure l'approche adoptée (*corpus-driven*) permet de « faire émerger de manière inductive des savoirs linguistiques » (Williams 2005: 13), concernant notamment le phénomène de la synonymie (diastratique et diatopique) en terminologie (« arcades » / « portiques » / « galerie » / « loge »). Au cours de cette expérience, les étudiant.e.s ont été ainsi sensibilisé.e.s aux variations terminologiques et discursives de la langue des Beaux-Arts et aux difficultés liées à l'identification des équivalents traductionnels les plus pertinents selon le niveau de spécialisation des textes (récits de voyage, textes littéraires, guides touristiques, textes de vulgarisation, manuels de critique et d'histoire de l'art, textes techniques), comme pour l'exemple que nous illustrerons relatif aux différentes dénominations des « portici » (it) en français (arcades, portiques, galeries, etc.) dans le domaine de l'architecture. Ils ont également appris à repérés dans les corpus des collocations spécialisées extraites des corpus LBC et qui ne sont pas prises en compte dans les ressources bi/plurilingues actuellement disponibles sur le marché (ex. dictionnaire bilingue français-italien Boch-Zanichelli), ainsi qu'à identifier de nouvelles stratégies pour la traduction de termes se référant à des réalités culturo-spécifiques (« portiques »). Voici quelques exemples significatifs, absents du Boch-Zanichelli (BZ):

- Collocations spécialisées de « colonne » (ex. « colonne engagée »), de « chapelle » (ex. « chapelle ardente, teintée ») ;
- Syntagme « mise au tombeau » PEINTURE et SCULPTURE : *deposizione* (ex. la mise au tombeau de Caravage, *la deposizione di Caravaggio*) – traduction de l'acception spécialisée absente du BZ.

La consultation systématique des corpus monolingues (FranText) et comparables dans les deux langues (LBC Français et LBC Italien, FrTenTen et ItTenTen) à des fins de comparaison s'est avérée fondamentale dans le processus d'apprentissage de la traduction spécialisée. Nous soulignerons en particulier les caractéristiques du corpus plurilingue *Lessico dei Beni Culturali* (LBC), un corpus spécialisé qui se distingue de par la qualité des données textuelles, ainsi que les limites, sur le plan qualitatif et quantitatif, d'autres grands corpus (Frantext,

FrenchTenTen) et d'un petit corpus récolté par les étudiants eux-mêmes à l'aide de l'outil BootCat.

Pour finir, les résultats obtenus dans les corpus ont été comparées avec les équivalents donnés par les principaux outils de TA disponibles gratuitement (Google Traduction, DeepL) qui se sont avérés tout à fait lacunaires, aussi bien dans la recherche des traductions des termes et des collocations spécialisées, prises singulièrement, qu'intégrés à des contextes plus larges puisées des corpus. Il s'avère que les traductions automatiques des collocations spécialisées se limitent à des traductions mot à mot des collocatifs. Ce constat a conduit les étudiant.e.s vers la prise de conscience du fait qu'il faut se méfier des outils de TA pour la traduction du domaine des Beaux-Arts et que le recours à des corpus spécialisés « de bonne qualité » s'avère incontournable.

En dernier ressort, les compétences acquises de la part des étudiant.e.s dans le cadre de cette expérience didactique sont multiples et ne se limitent pas à des compétences purement linguistiques : elles relèvent de l'apprentissage de la méthodologie de la recherche (récolte de données + initiation à leur analyse), de compétences technologiques (emploi de logiciels pour différentes tâches, BootCat, Atlas.ti, SketchEngine) et de competences professionalissantes pour les métiers du traducteur, du terminographe, du documentaliste numérique et aussi d'opérateur touristique plurilingue.

Références bibliographiques

Boulton, A., Tyne, H. (2014). *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues.* Paris: Didier.

Dechamps, Ch. (2020). Compétence stratégique et corpus : quelques pistes pour la formation en traduction. In M. Célio Conceição & M. T. Zanola (Eds.), *Terminologia e mediação linguística: métodos, práticas e atividades*, (pp. 107-18). Faro: Universidade do Algarve Editora.

Dechamps, Ch. (2023). Glossaire terminologique collaboratif et 'Data-Driven Learning' dans le cadre de la traduction du lexique artistique, In V. Zotti & M. Turci (Eds.), *Nuove strategie per la traduzione del lessico artistico. Da Giorgio Vasari a un corpus plurilingue dei beni culturali* (pp. 129-142). Firenze: Firenze University Press.

Farina, A., Sini, L. (2020). Il corpus LBC francese. In R. Billero, A, Farina & M. C. Nicolás Martínez (Eds.), *I corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali* (pp. 77-100). Firenze: Firenze University Press.

Gandin, S. (2016). Translating and Learning the Language of Tourism as LSP: Corpusbased Approaches. In G. Garzone, D. Heaney & G. Riboni, *Focus on LSP Teaching: Developments and Issues*, (pp. 65-82). Milano: LED.

Hurtado Albir, A. (2008). Compétence en traduction et formation par compétences. *TTR Traduction, Terminologie, Rédaction* 21(1), 17-64.

Kilgarriff, A. et al. (2014). The Sketch Engine: ten years on. *Lexicography* 1, 7-36.

L'Homme, M. (2004). *La terminologie : principes et techniques.* Montréal : Presses de l'Université de Montréal.

Lino, M. T. (2001). De la néologie à la lexicographie spécialisée d'apprentissage. *Cahiers de Lexicologie* 78(1), 139-45.

Turci, M., Aragrande, G. (2020). On translating art and heritage discourse from Italian into English: From a learner corpus to a specialized corpus. In. A. P. Alamán & V. Zotti (Eds.), *The Language of Art and Cultural Heritage* (pp. 12-38). Newcastle: Cambridge Scholars Publishing.

Zotti, V. (2023). Traduire en français le lexique du patrimoine de la ville de Bologne : le sous-corpus BER du projet LBC. In V. Zotti & M. Turci (Eds.), *Nuove strategie per la traduzione del lessico artistico. Da Giorgio Vasari a un corpus plurilingue dei beni culturali* (pp. 191-223). Firenze: Firenze University Press.