JOURNÉES de LINGUISTIQUE de CORPUS

21-23 octobre 2025

ENS de Lyon

icar

PRAXILING

LIDILEM Université Grenoble Alpes

CLLE COGNITION LANGUES LANGAGE ERGONOMIE

DDL LABORATOIRE DYNAMIQUE DU LANGAGE

ENS DE LYON

cnrs

UNIVERSITÉ LUMIÈRE LYON 2

'Apparemment elle cracherait du sang' : Évidentialité et co-construction du doute lors du transfert d'appel d'urgence

Marine Riou[1,2]
[1] Centre de Recherche en Linguistique Appliquée (CeRLA), Université Lumière Lyon 2
[2] Institut Universitaire de France (IUF)
marine.riou@univ-lyon2.fr

Introduction

L'évidentialité correspond aux moyens linguistiques par lesquels le locuteur exprime la source d'une information (Aikhenvald, 2004), par exemple pour signaler qu'elle est issue d'une perception, par inférence, ouï-dire, etc. Avec l'évidentialité reportative, le locuteur signale qu'il n'est pas la source de l'information, qui est donc de seconde main – ou « *type 2 knowledge* » selon Pomerantz (1980). Bien que les recherches sur l'évidentialité soient très riches et nombreuses, ce domaine a peu été étudié avec des données interactionnelles (Clift, 2006; Fox, 2001; Mori et al., 2017; Mushin, 2001), ce qu'on peut contraster à l'intérêt soutenu pour l'étude de l'épistémicité en interaction (Heritage, 2013 *inter alia*). Une raison possible à cette disparité est qu'une grande partie des recherches sur l'interaction porte sur des langues dans lesquelles l'expression de l'évidentialité n'est pas obligatoire et pas ou peu grammaticalisée. Pour le français, on peut citer entre autres marqueurs évidentiels *visiblement* et *faut croire* (inférence), *à ce qu'on dit* et *dit-on* (ouï-dire) (Dendale, 2022), ainsi que le conditionnel dit « épistémique » (Kronning, 2012) ou « de reprise » (Dendale & Coltier, 2012). Cette étude propose de se pencher sur un rare contexte interactionnel dans lequel le marquage évidentiel est une pratique interactionnelle fréquente en français oral.

Corpus

Les données proviennent d'un corpus de 200 appels téléphoniques reçus par un Service d'Aide Médicale Urgence (SAMU) entre 2018 et 2022. La collection créée pour cette étude contient 151 occurrences de marqueurs d'évidentialité reportative, répartis dans 104 tours de parole. Seuls les enregistrements audios des transferts d'appel entre différents acteurs institutionnels ont été pris en compte. Il s'agit soit d'un transfert entre les sapeurs-pompiers et le SAMU, soit d'un transfert au sein du SAMU entre un assistant de régulation médicale (ARM) et un médecin régulateur. Dans les deux cas, l'appelant est mis en attente pendant qu'une brève transmission orale a lieu entre deux représentants institutionnels, comme l'illustrent les extraits présentés (figure. 1, figure. 2).

```
   1   MEDECIN  allo ?
   2   ARM      hh oui ((NOM)) tu prends j- euh tu prends la fiche ((NUMERO)) sur la
   3            gauche. .h c'est un voisin qui appelle, pour u:n monsieur de: soixante
   4            quinze ans,
   5            (0.2)
→  6   ARM      .h [eu  ]:h qui aurai:t apparemment une oppression à la poitrine ?
   7   MEDECIN     [oui,]
```

figure . 1 **Extrait d'une transmission entre assistant de régulation médicale (ARM) et médecin régulateur**

```
1    POMPIER  une demande de régulation s'il te plait, à ((COMMUNE)).
2             (0.5)
3    ARM      à ((COMMUNE)) oui:, hh madame ((NOM)) ?
4             c'est ça ?
5             quatre [vingts ans ? ]
6    POMPIER         [°voilà°      ] c'est sa voisine qui appelle,
→ 7           apparemment elle cracherait du ↑sang cette dame ?
8             .hhh [(inaud.) sa voisine] eu:h
9    ARM           [°d'a:cco:rd°       ]
10   POMPIER   elle ne sait pas du tout euh quels antécédents il a ?
```

figure . 2 Transmission entre sapeur-pompier et ARM

Méthodologie

Cette étude s'inscrit dans le cadre de la linguistique interactionnelle. Les enregistrements audio ont été transcrits de manière détaillée selon les conventions usuelles de l'analyse conversationnelle adaptées pour le français (ICAR UMR 5191, 2013). L'immersion non-motivée dans les données a permis d'identifier le phénomène d'intérêt puis la création d'une collection de cas. Chaque cas a fait l'objet d'une analyse *in situ*. Une attention particulière a été portée au devenir des informations accompagnées de marqueurs d'évidentialité reportative, pour déterminer leur possible impact sur la trajectoire interactionnelle et institutionnelle des appels, plus particulièrement dans l'interaction suivante lorsque le second régulateur reprend l'appelant au téléphone.

Résultats

Lors du transfert d'un appel d'urgence, le marquage évidentiel est une pratique interactionnelle couramment déployée pour indiquer qu'une information provient de l'appelant. Les régulateurs sapeurs-pompiers et du SAMU mobilisent tout particulièrement le discours indirect (*elle me dit que*), l'adverbe *apparemment*, et le conditionnel de reprise (*il serait inconscient*).

Les données montrent l'orientation des régulateurs vis-à-vis des marqueurs évidentiels lors des transmissions, et leurs efforts interactionnels conséquents pour caractériser finement le niveau de certitude ou de doute qui perdure. Ce travail évidentiel s'explique par un contexte interactionnel où il est particulièrement difficile de coconstruire le savoir alors que certaines informations sont cruciales pour la prise en charge médicale. Cependant, le recours aux marqueurs d'évidentialité reportative est à double tranchant et peu avoir des effets pervers.

L'analyse de la suite des appels (une fois la transmission réalisée) montre que dans ce contexte institutionnel, l'évidentialité reportative va au-delà de la suspension de la prise en charge, pour mettre en doute la validité de l'énoncé, et signaler qu'il s'agit d'une information demandant confirmation, comme dans les emplois journalistiques du conditionnel de reprise (Gosselin, 2001). Or, *toutes* les informations transmises entre régulateurs sont par définition issues d'une autre source, à savoir l'appelant. Il est ainsi clair et évident pour les participants impliqués dans l'activité de transmission qu'il s'agit d'informations rapportées. Souligner explicitement qu'un élément est issu d'une autre source attire donc l'attention sur la non-prise en charge énonciative, à savoir le refus de se positionner sur la validité du dictum. Le marquage évidentiel est alors interprété par les participants comme signalant une précaution toute particulière, et donc une fausse neutralité quant à la validité de la proposition.

Cette co-construction du doute peut bénéficier aux patients quand le régulateur suspecte que la situation est plus préoccupante que ce que les paroles de l'appelant peuvent laisser croire. Cependant, l'évidentialité reportative a aussi pour effet potentiel de présenter l'appelant comme une source peu fiable ou insuffisamment compétente et informée. Ceci peut conduire à des

délais inutiles (par exemple si le second régulateur vérifie à nouveau une information pourtant déjà établie) ou un désalignement voire un conflit avec l'appelant. Dans les cas les plus graves, la non-prise en charge énonciative peut avoir un impact sur la prise en charge médicale.

Références bibliographiques

Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford University Press.

Clift, R. (2006). Indexing stance: Reported speech as an interactional evidential. *Journal of Sociolinguistics*, 10(5), 569–595.

Dendale, P. (2022). Evidentiality in French. In B. Wiemer & J. I. Marin-Arrese (Eds.), *Evidential Marking in European Languages* (pp. 171–234). De Gruyter Mouton.

Dendale, P., & Coltier, D. (2012). La lente reconnaissance du "conditionnel de reprise" par les grammaires du français. In B. Colombat, J.-M. Fournier, & V. Raby (Eds.), *Vers une histoire générale de la grammaire française. Matériaux et perspectives* (pp. 631–652). Champion.

Fox, B. A. (2001). Evidentiality: Authority, responsibility, and entitlement in English conversation. *Journal of Linguistic Anthropology*, 11(2), 167–192.

Gosselin, L. (2001). Relations temporelles et modales dans le conditionnel journalistique. In P. Dendale & L. Tasmowski (Eds.), *Le conditionnel en français* (pp. 45–66). Klincksieck.

Heritage, J. (2013). Epistemics in conversation. In J. Sidnell & T. Stivers (Eds.), *The Handbook of Conversation Analysis* (pp. 370–394). Wiley-Blackwell.

ICAR UMR 5191. (2013). *Convention ICOR*. CNRS ; Lyon 2 ; ENS de Lyon.

Kronning, H. (2012). Le conditionnel épistémique : Propriétés et fonctions discursives. Langue Française, 173(1), 83–97.

Mori, J., Imamura, A., & Shima, C. (2017). Epistemic management in the material world of workplace: A study of nursing shift handovers at a Japanese Geriatric Healthcare Facility. *Journal of Pragmatics*, 109, 64–81.

Mushin, I. (2001). Japanese reportive evidentiality and the pragmatics of retelling. *Journal of Pragmatics*, 33(9), 1361–1390.

Pomerantz, A. (1980). Telling my side: "Limited access' as a "fishing" device. *Sociological Inquiry*, 50(3–4), 186–198.

« en fait… du coup… voilà(…) quoi » : quelques marqueurs discursifs dans le discours de soignants

Audrey Roig[1]
EDA (URP 4071), Université Paris Cité
audrey.roig@parisdescartes.fr

À partir de l'analyse syntaxique et sémantique de paroles de soignants recueillies au terme de la 1re année de covid-19 (cf. description du corpus *infra*), cette communication vise, en première intention, à examiner la distribution de différents marqueurs, à savoir : *en fait, du coup, voilà* et *quoi.*

La distribution ainsi que les pourcentages d'apparition de ces marqueurs dans les discours des soignants seront alors commentés à l'aune de données sociologiques telles que

- l'âge du locuteur (avec/sans la présence d'adolescents ou jeunes adultes au sein du foyer, ce qui pourrait possiblement influencer le mode d'expression du locuteur),
- la fonction hospitalière de l'enquêté (aide-soignant, infirmier, interne, titulaire), ou encore
- son service d'appartenance (odontologie, réanimation, urgences).

Voici, par exemple, les résultats liminaires obtenus selon les données liées à la fonction hospitalière des soignants :

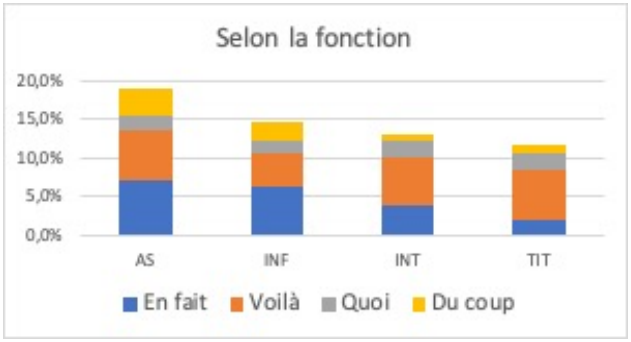


figure . 1 Répartition des marqueurs *en fait, voilà, quoi, du coup* selon la fonction hospitalière

En seconde intention, cette communication se focalisera exclusivement sur les locutions *du coup*. Nous discuterons notamment la portée de *du coup*, et soulèverons quelques problèmes que l'identification de la frontière entre les énoncés à l'oral pose pour l'analyse d'une telle forme. Nous nous attarderons aussi sur les conditions d'apparition de cette forme dans le discours. L'observation de la figure 1 montre que son utilisation varie selon les fonctions des agents hospitaliers : peu utilisée par les titulaires et par les internes, la locution *du coup* est davantage remarquée dans les discours des infirmiers et, surtout, des aides-soignants. Y a-t-il une explication à ce phénomène ? Des éléments de réponse à cette question seront apportés par une analyse du corpus.

Le corpus qui sert à cette analyse est le fruit du projet de recherche AS2-HP. Il est constitué de 33 entretiens filmés et transcrits, d'1 heure chacun environ. Constitué entre mai et juin 2021, il rassemble les paroles d'aides-soignants, d'infirmiers, d'internes et de titulaires de différents

services hospitaliers (odontologie, urgences, réanimation) de plusieurs établissements parisiens de l'AP-HP. Les discours rassemblés dans ce corpus portent globalement sur la façon dont ces soignants ont vécu la première année de pandémie de covid-19, sur les émotions qu'ils ont ressenties face aux difficultés sociales et matérielles que cette pandémie a engendrées.

Bibliographie indicative

Abouda Lofti, 2022, « L'émergence du marqueur méta-discursif *du coup* : de la conséquence à l'actualisation énonciative », *Langages*, 226 : 99-116.

Andersen Hanne Leth, 2007, « Marqueurs discursifs propositionnels », *Langue française*, 154 : 13-28.

Beeching Kate, 2007, « La co-variation des marqueurs discursifs *bon, c'est-à-dire, enfin, hein, quand même, quoi* et *si vous voulez* : une question d'identité ? », *Langue française*, 154 : 78-93.

Chanet Catherine, 2001, « 1700 occurrences de la particule *quoi* en français parlé contemporain : approche de la "distribution" et des fonctions en discours », *Marges Linguistiques*, 2 : 56-80.

Col Gilles, Danino Charlotte, Bikialo Stéphane (dir.), 2020, *Polysémie, usages et fonctions de « voilà »*, Berlin, De Gruyter.

Col Gilles & Knutsen Dominique, 2024, « Validation psychologique d'hypothèses linguistiques, et inversement. Autour de l'expression "du coup" », *Essais*, 21 : en ligne.

Dostie Gaétane, 2004, *Pragmaticalisation et marqueurs discursifs : analyse sémantique et traitement lexicographique*, Bruxelles, De Boeck/Duculot.

Foucher Stenkløv Nelly, 2015, « *Du coup* : un connecteur plus ou moins "logique" de l'argumentation orale », *Synergies Pays Scandinaves*, 10 : 11-22.

Malm Katerine, 2011, *Une étude de l'expression adverbiale « du coup »*, Mémoire de Master de l'Université de Tromsø : en ligne.

Nielsen Marina, *La polysémie et le mot* coup, Åbo, Åbo Akademi University Press, 2004.

Paillard Denis, 2021, *Grammaire discursive du français. Étude des marqueurs discursifs en -ment*, Bruxelles, PIE Peter Lang.

Roig Audrey, 2018, *Complexité phrastique : construction, linéarisation, marquage*, mémoire d'HDR, Sorbonne Université.

Rossari Corinne, Jayez Jacques, 2000, « *Du coup* et les connecteurs de conséquence dans une perspective dynamique », *Lingvisticæ Investigationes*, 23/2 : 303-326.

"Pas vrai ?" : diversité des usages, des fonctions et évolution

Laurence Buson[1], Vannina Goossens[1], Jean-Pierre Chevrot[1], Aurélie Nardy[1]
[1] Laboratoire Lidilem, Université Grenoble Alpes
laurence.buson@univ-grenoble-alpes.fr, vannina.goossens@univ-grenoble-alpes.fr,
jean-pierre.chevrot@univ-grenoble-alpes.fr, aurelie.nardy@univ-grenoble-alpes.fr

Introduction et cadrage

Les interrogatives en français sont caractérisées par une multiplicité de formes et de fonctions (Badin et al., 2021 ; Coveney, 2020 ; Gadet, 2020 ; Hansen, 2001 ; Larrivée & Guryev, 2021 ; Quillard, 2001; Reinhardt, 2021). Cette diversité de formes chez l'adulte se retrouve précocement chez l'enfant (Buson et al., à paraitre ; Gillet & Benzitoun, 2024). Une forme d'interrogative totale attestée chez les enfants en maternelle se révèle en revanche extrêmement rare chez l'adulte (Buson et al., à paraitre) : il s'agit de l'énoncé averbal préfabriqué (Lefeuvre, 2020b, 2020a ; Tutin, 2022), renforçateur de type *tag* (Druetta, 2009 ; Guryev, 2017), "pas vrai ?". De "on est pas des bébés, pas vrai ?" en maternelle, au "pas vrai @X ?" des réseaux sociaux, en passant par l'oral représenté dans les dialogues de romans, cet énoncé préfabriqué des interactions nous est familier. On le trouve dans des dialogues de romans déjà au XVIII^{e} siècle. Aujourd'hui, il est fréquent dans les tweets. Cependant, son absence dans plusieurs corpus oraux d'adultes interroge : s'agit-il d'une forme essentiellement enfantine ? Sa connotation familière et/ou populaire donne-t-elle à voir une représentation de l'oral plutôt que l'oral lui-même ? Sa distance ironique et son potentiel d'interpellation polémique expliquent-ils son succès sur les réseaux sociaux ? Au final, est-ce une forme dont les usages adultes sont principalement stylisés, ou une simple variante familière de "n'est-ce pas" ?

Après avoir décrit la structure de cette séquence, nous nous intéresserons à sa fréquence et sa répartition dans différents corpus, oraux et écrits, d'enfants et d'adultes, avant d'analyser d'une manière plus qualitative les différentes valeurs pragmatiques de cette forme, entre demande de confirmation/d'adhésion/d'approbation et provocation. Nous interrogerons l'hypothèse d'une revitalisation de cette forme, en lien avec ses valeurs de connivence et de polémique. Nous nous demanderons également si la potentielle revitalisation de cette structure ne se ferait pas en écho à la disparition progressive de ce qui est souvent présenté comme son équivalent formel, "n'est-ce pas" (Combettes, 2016).

Corpus et méthodologie

Pour étudier la séquence "pas vrai", dont la fréquence d'apparition est très variable en fonction des contextes interactionnels, nous avons rassemblé un ensemble volumineux de corpus diversifiés. Ceux-ci nous permettront d'étudier le fonctionnement de "pas vrai" dans des interactions impliquant des adultes et/ou des enfants, à l'oral ainsi qu'à l'écrit.

Nos analyses portent sur quatorze corpus contemporains : 6 corpus de parole ou d'interactions orales transcrites, et 8 corpus d'écrits ou d'écrits oralisés. Nous nous appuyons notamment sur

les corpus rassemblés et retraités dans le cadre du projet ANR Prefab[1] (piloté par A. Tutin) pour l'étude des phrases préfabriqués des interactions.

Les corpus oraux nous donnent accès à de l'oral spontané (parole enfantine dans le corpus *DyLNet*, « sms vocaux » dans le corpus *Les vocaux*) ainsi qu'à des interactions entre adultes ou adulte-enfant, notamment issues d'entretiens (corpus *MPF, ESLO2, ORFEO-CEFC, TCOF2*).

Les corpus écrits sont également diversifiés : écrits oralisés sous forme de dialogues de films (corpus *Miscellanées*) et de dialogues de dessin animé (corpus *Maya l'abeille*) ; articles de presse (corpus d'échantillons d'articles du *Monde*) ; dialogues de romans français contemporains (corpus créé dans l'ANR-DFG PhraseoRom et retraité dans l'ANR Prefab, notamment pour baliser automatiquement le discours direct) ; interactions écrites (corpus *Wikidiscussions*, *SoSweet*) ; contributions à des débats en ligne (corpus le *Vrai débat*, le *Grand débat*). Ces corpus seront analysés en prenant en compte leurs différents genres textuels à l'aide des plateformes Lexicoscope 2.0 (Kraif & Diwersy, 2012) et TXM (Heiden et al., 2010). Notre approche combinera des analyses quantitatives à l'examen qualitatif des contextes d'usage. Le tableau 1 ci-après recense les caractéristiques des corpus utilisés.

Premiers résultats

Les premières analyses indiquent que les énoncés préfabriqués interrogatifs en "pas vrai ?" sont attestés chez les enfants de maternelle, ce qui pourrait s'expliquer par sa simplicité de construction, à l'image des interrogatives par intonation, qui sont de loin les plus représentées à cet âge (Buson et al., à paraitre). En revanche, nous montrerons que cette structure est quasiment absente des corpus oraux adultes ; quand elle est utilisée, elle est associée à des stylisations. Cette forme est par ailleurs présente dans l'oral représenté, comme les dialogues de romans, dans lesquels nous analyserons les valeurs socio-stylistiques qui leur sont le plus souvent associées (locuteurs de milieux populaires, registres familiers, en association avec d'autres marques, comme les trucages orthographiques). Et elle est surtout massivement présente dans les interactions informelles en ligne comme les tweets, à raison de milliers d'occurrences.

Ces résultats, en association avec une analyse fonctionnelle, nous amèneront à interroger l'hypothèse d'une revitalisation de cette forme chez l'adulte, par le biais de ces usages en ligne. En effet, la locution "pas vrai" permet une interpellation facile de l'interlocuteur ou d'un témoin, notamment quand elle est suivie d'un "@" ou d'un "#", que ce soit dans un but de recherche de connivence et d'adhésion, ou d'interpellation polémique. Ces différentes analyses, (socio)linguistiques et pragmatiques, quantitatives et qualitatives, apporteront des éléments de compréhension sur les usages de "pas vrai" chez différents types de locuteurs et sur son évolution.

1 https://lidilem.univ-grenoble-alpes.fr/node/16/axes-recherche/axe-1-descriptions-linguistiques-tal-corpus/projets-axe-1/constructions-phrases-prefabriquees-interactions-langagieres-prefab-anr-2022-2026

Corpus	Type de données	Taille du corpus
Parole ou interactions orales authentiques		
DyLNet[2]	Productions orales non supervisées d'enfants de 3 à 6 ans	674 795 mots transcrits
ORFEO-CEFC[3] [8]	Interview conversationnelles, interactions en situations naturelles, entretiens etc.	3 510 300 tokens
MPF[4] [8]	Entretiens et événements sans enquêteur	977 039 tokens
ESLO2[5] [8]	Entretiens, conférences, enregistrements lors de repas etc.	1 670 575 tokens
TCOF2[6] [8]	Entretiens, conversations, réunions etc.	426 729 tokens
Les vocaux[7] [8]	« SMS vocaux »	257 870 tokens
Ecrit oralisé		
Maya l'abeille[8]	Dialogues de dessins animés	141 interrogatives 1094 tokens
Miscellanées[9]	Sous-titres TV	9 867 437 tokens
Corpus presse et romans		
PhraseoRom romans (post 1980)[8] [10]	890 romans français de 1980 à 2016 avec balisage du discours direct	83 952 614 tokens
Presse (2010-2019)[8]	Échantillons d'articles du *Monde*	70 372 487 tokens
Interactions écrites		
Wikidiscussions[8] [11]	Discussions lors de la (co)rédaction d'articles Wikipédia en français (extrait d'EFG WikiCorpus)	60 290 590 tokens
Sosweet[12]	Tweets produits entre 2006 et 2019 par 2 878 562 utilisateurs	658 747 413 tweets
Grand débat[8]	Contributions au Grand débat national de 2019	179 895 852 tokens
Vrai débat [8]	Contributions à la plateforme de revendications en ligne créée par un groupe de gilets jaunes	2 168 091 tokens

table 1. : Corpus utilisés pour l'analyse

2 Nardy, A., Bouchet, H., Rousset, I., Liégeois, L., Buson, L., Dugua, C., & Chevrot, J.-P. (2021). Variation sociolinguistique et réseau social : Constitution et traitement d'un corpus de données orales massives. Corpus, 22. https://doi-org.sid2nomade-2.grenet.fr/10.4000/corpus.5561.

3 Benzitoun, C., Debaisieux, J. M. (Eds) (2020). Orféo: un corpus et une plateforme pour l'étude du français contemporain. Langages, 219(3).

4 Gadet F. (dir.) (2017). Les parlers jeunes dans l'Ile-de-France multiculturelle. Ophrys.

5 Abouda, L., Baude, O. (2006). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO. In Rastier, F., Ballabriga, M. (Eds.). Corpus en lettres et sciences sociales : des documents numériques à l'interprétation (pp. 143-50). Texto.

6 ATILF (2024). TCOF : Traitement de Corpus Oraux en Français [Corpus]. ORTOLANG, www.ortolang.fr, v2.2, https://hdl.handle.net/11403/tcof/v2.2.

7 Glikman, J., Mazziotta, N., Benzitoun, C. et al. (2025). LesVocaux [Corpus]. ORTOLANG, https://hdl.handle.net/11403/lesvocaux/v0.0.2.

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*. https://doi-org.sid2nomade-2.grenet.fr/10.4000/corpus.5752

Buson, L., Chevrot, J.-P., Nardy, A., & Rousset, I. (à paraitre). Acquisition of syntactic variation : Production of interrogatives in French preschool children. In L. Rosseel & E. Zenne (Éds.), *Acquiring Language variation : The interaction between world and mind*. John Benjamins.

Combettes, B. (2016). La disparition d'un marqueur discursif : Le cas de n'est-ce pas. *Journal of French Language Studies*, *26*(1), 13-28. https://doi.org/10.1017/S0959269515000460

Coveney, A. (2020). L'interrogation directe. In *Encyclopédie grammaticale du français*. http://www.encyclogram.fr/notx/002/002_Notice.php

Druetta, R. (2009). *La question en français parlé : Étude distributionnelle*. Trauben Edizioni.

Gadet, F. (2020). Langue et variation. In *Encyclopédie Grammaticale du Français*. http://www.encyclogram.fr/notx/019/019_Notice.php#tit4

Gillet, P., & Benzitoun, C. (2024). Quelle est la grammaire des enfants ? L'exemple des interrogatives partielles en français. *Linx [online]*, *87*. https://doi-org.sid2nomade-2.grenet.fr/10.4000/12zrj

Guryev, A. (2017). *La forme des interrogatives dans le Corpus suisse de SMS en français : Étude multidimensionnelle* [Thèse de doctorat]. Université de Neuchâtel.

Hansen, M.-B. M. (2001). Syntax in interaction : Form and function of yes/no interrogatives in spoken standard French. *Studies in Language*, *25*, 463-520.

Heiden, S., Magué, J.-P., & 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*, 1021-1032. http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf

8 Universität Konstanz (2016). Corpus oral de questions (Maya l'Abeille) [Corpus]. ORTOLANG, https://hdl.handle.net/11403/maya/v1.

9 Corpus disponibles sur le Lexicoscope 2.0 : http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/.

10 Diwersy, S., Gonon, L., Goossens, V., Kraif, O., Novakova, I., Sorba, J. et Vidotto, I. (2021). La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase, Corpus [En ligne], 22, https://doi.org/10.4000/corpus.6101.

11 Mai Ho-Dac, L. (2024). EFG_WikiCorpus - discussions en coulisse de Wikipedia (anglais, français, allemand) [Corpus]. ORTOLANG, https://hdl.handle.net/11403/efg-wikicorpus/v1

12 ICAR, DANTE Inria, LIDILEM, ALMANACH (2024). SoSweet [Corpus]. ORTOLANG, https://hdl.handle.net/11403/sosweet/v1.

Kraif, O., & Diwersy, S. (2012). Le Lexicoscope : Un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques. *Actes de la conférence TALN 2012*, 399-406.

Larrivée, P., & Guryev, A. (2021). Variantes formelles de l'interrogation. Présentation. *Langue française*, *212*(4), 9-24. https://doi.org/10.3917/lf.212.0009

Lefeuvre, F. (2020a). Les marqueurs discursifs averbaux résomptifs. In F. Diémoz & et al (Éds.), *Le Français innovant* (p. 225-243). Peter Lang. halshs- 03143412

Lefeuvre, F. (2020b). Vrai comme marqueur discursif. In M. Saiz-Sánchez & et al. (Éds.), *Marques d'oralité et représentation de l'oralité en français* (p. 127-148). Presses Universitaires Savoie Mont Blanc. halshs-03143458

Quillard, V. (2001). La diversité des formes interrogatives : Comment l'interpréter ? *Langage et société*, *95*(1), 57-72. https://doi.org/10.3917/ls.095.0057

Reinhardt, J. (2021). L'intonation des interrogatives par maintien de l'ordre SV. *Langue française*, *212*(4), 41-56. https://doi.org/10.3917/lf.212.0041

Tutin, A. (2022). Comment dirais-je ? Que veux-tu ? Comment ça va ? Quelques observations sur les phrases interrogatives partielles préfabriquées dans les interactions orales et les dialogues romanesques. *Lingvisticae investigationes: International Journal of Linguistics and Language*, *45*(2), 172-196. https://doi.org/10.1075/li.00072.tut

“N’importe quoi !”, jette-t-elle enfin… Introducteurs de discours directs et phrases préfabriquées des dialogues

Vannina Goossens, Agnès Tutin, Adam Renwick
Laboratoire LIDILEM, Université Grenoble Alpes
vannina.goossens@univ-grenoble-alpes.fr ; agnes.tutin@univ-grenoble-alpes.fr ;
adam.renwick@univ-grenoble-alpes.fr

Introduction

Les dialogues romanesques regorgent de phrases préfabriquées stéréotypées qui stylisent l’oral spontané, comme dans les exemples suivants (Tutin & Gharbi, 2020) :

- Djeeb répondit aux questions de bonne grâce alors que Sagace jetait sur ce qui l’entourait des regards d’enfant rêveur. Il ne sembla s’éveiller que lorsque le sujet d’autres tâches fut abordé.
  - Pas question ! clama-t-il avec un regain de fierté bravache.
  Gidon, Laurent (2009). *Djeeb le chanceur*.

- Elle effleura son espace, elle aurait pu lire l’ordonnance, alors il posa la main dessus comme une gifle, il la désintégra de son regard, elle battit en retraite. « Mais **ce n’est pas possible** ! hurla-t-il.C’est toujours comme ça !Ils ne restent jamais à leur place !Toujours à resquiller !Il faut avoir les yeux dans le dos ! »
  Jenni, Alexis (2011). *L’art français de la guerre*.

Ces phrases préfabriquées sont des énoncés autonomes et récurrents, associés à un acte de langage, particulièrement fréquents dans les interactions. Dans les dialogues romanesques, les séquences comportant ces phrases préfabriquées sont souvent introduites par des verbes de parole, *clamer* et *hurler* dans les deux exemples ci-dessus, parfois associés à des modifieurs comme *avec un regain de fierté bravache* dans le premier exemple.

Cette communication se propose d’examiner la façon dont sont combinés ces deux types d’éléments dans les séquences dialoguées, afin d’observer comment est stylisé l’oral spontané dans le roman, qui ne reproduit bien entendu pas l’oral réel (Durrer, 1990). Nous souhaitons ainsi observer la cooccurrence entre introducteurs de dialogues et phrases préfabriquées en répondant à plusieurs questions. On essaiera dans un premier temps de repérer la diversité des introducteurs en observant les différents types d’actes de langage associés. Apparaissent-ils prévisibles selon la fonction de la phrase ? Par exemple, une phrase fortement expressive comme *c’est pas possible* est-elle davantage corrélée à des introducteurs du même type (comme *hurler* dans le second exemple), renforçant ainsi la valeur expressive ? Ou au contraire, l’introducteur sert-il à moduler le sens des expressions, en particulier lorsqu’elles présentent une forme d’ambiguïté ?

Phrases préfabriquées et introducteurs de discours directs

Les séquences introductrices de discours rapporté jouent, dans les romans, un rôle qui dépasse le seul enjeu de lisibilité des dialogues et des normes stylistiques pèsent sur leur emplois (Salvan, 2005). Elles sont ainsi un poste d'observation intéressant pour étudier le style des auteurs, ceux-ci pouvant faire le choix de se passer quasi totalement de telles marques ou, au contraire, en faire un usage systématique, mais aussi pour s'intéresser aux caractéristiques des sous-genres romanesques, notamment en ce qui concerne la littérature populaire (Diwersy et al., 2020). Les verbes, dans ces séquences introductives, ne se limitent pas aux seuls verbes de parole (« un verbe de parole est un verbe qui dénote une activité linguistique du sujet parlant visant normalement à communiquer un message à quelqu'un », Lamiroy & Charolles,2008) et l'on retrouve à leur côté des verbes psychologiques (*penser*), des verbes de sentiment (*s'affoler*) ou encore des verbes de mimiques (*grimacer*). Parmi les typologies existantes (par exemple Lamiroy & Charolles, 2008 ; Rosier, 2008 ; Anscombre, 2015) nous nous appuierons sur celle de Harras et al. (2004) qui présente l'intérêt de classer les verbes de parole en fonction des actes de langage qui leur sont associés.

Corpus et méthodologie

Corpus

Le corpus exploré comporte un ensemble de romans contemporains en français (1980-2020), développé dans le cadre du projet PhraseoRom (Novakova & Siepmann, 2019), de 83,9 millions de tokens, comprenant de nombreux dialogues. Les parties dialoguées ont été isolées, ce qui a permis des recherches ciblées sur ces séquences spécifiques. Le corpus se compose de 1123 romans répartis en 6 sous-genres : romans de littérature générale (GEN), sentimentaux (SENT), policiers (POL), historiques (HIST), de science-fiction (SF) et fantasy (FY).

Méthodologie

Une première extraction d'un sous-ensemble de phrases préfabriquées a été effectuée en exploitant l'outil Lexicoscope (Kraif, 2019). Le corpus étant analysé syntaxiquement avec Stanza (Qi et al., 2020) nous avons pu utiliser les relations de dépendances syntaxiques pour construire des requêtes ciblées. Nous avons choisi d'étudier 10 phrases préfabriquées fréquentes et diversifiées en recherchant, dans le discours direct uniquement, les séquences lexico-syntaxiques suivantes, lorsqu'elles sont suivies d'une ponctuation : *du calme* (165 occ.), *n'importe quoi* (399 occ.), *pas possible* (419 : occ.), *pas question* (215 occ.), *qu'est-ce qu'il se passe* (416 occ.), *que veux-tu dire* (471 occ.), *si tu veux* (575 occ.), *ta gueule* (239 occ.), *tu te rends compte* (289 occ.), *pas vrai* (89 occ.).

Les contextes de ces PPI sont par la suite dépouillés manuellement, d'une part pour exclure les séquences qui ne relèvent pas d'une PPI et, d'autre part, pour identifier la présence ou non d'un verbe introduisant le tour de parole dans lequel se trouve la PPI.

Une grille d'analyse a été développée intégrant différents paramètres :

- Le type de verbe introducteur de discours : verbe de parole (*dire*, *crier*…), verbe psychologique (*penser*, *imaginer*…), autre verbe (*sourire*, *bailler*…).
- La classe sémantique du verbe introducteur, en s'appuyant sur la typologie de Harras et al. (2004) : verbes génériques (*dire*, *parler*…), expressifs (*critiquer*, *saluer*…), structurant l'interaction (*répondre*, *ajouter*…), etc.
- La position du verbe par rapport au discours rapporté : antéposé ([…] *il dit : « Hé ! du calme,* […] *»*) ou postposé (*« Ta gueule ! » aboya-t-il.*).

- La portée du verbe par rapport au tour de parole et la PPI : à proximité immédiate de la PPI comme dans les exemples ci-dessus ou portant sur un tour de parole plus vaste (*Je vais vous dire, lui annonça-t-il sur le ton de la confidence… Je serais prêt à parier que votre venue ici va changer bien des choses ! Pas vrai, Rosette ?*).
- La présence d'une construction particulière faisant partie intégrante de la séquence introductive de discours : syntagme prépositionnel (*dit Yacine d'une voix monocorde qui contrastait violemment avec la colère qui giclait de ses prunelles* ; *ai-je dit avec plus de véhémence que je ne l'aurais voulu*), structure avec gérondif (*s'enthousiasme Jobelot en s'adressant à la jeune fille qui l'a rejoint dans le courant de la rivière*), adverbe de manière (*ajoute-t'il vivement*), etc.

Premiers résultats

Les premiers codages font apparaître des différences de fonctionnement des PPI par rapport aux introducteurs de discours direct. Par exemple, *pas question* et *ta gueule*, qui sont toutes deux des PPI expressives, ne sont pas associées aux mêmes types de verbes : *pas question* est le plus souvent associée à des verbes qui structurent l'interaction (*riposter*, *rétorquer*, *couper*, *objecter*...) alors que *ta gueule* est le plus souvent associée à des verbes qui expriment le mode d'expression (*crier*, *hurler*, *beugler*, *grogner*...). Il sera nécessaire d'affiner ce premier niveau d'analyse des verbes, *démarrer* ou *finir* par exemple ne structurant pas l'interaction de la même manière que *couper* qui a une fonction bien spécifique. *Du calme*, PPI également expressive, privilégiera pour sa part les verbes génériques (*dire* et *faire*).

La proportion de PPI dans un tour de parole introduit par un verbe introducteur est également très variable : par exemple, 54% des occurrences de *du calme* sont introduites par un verbe, contre 37% des *ta gueule*, 29% pour *pas vrai* et 12% des *tu te rends compte*. Parmi ces occurrences 100% des verbes introduisant *du calme* se situent à proximité immédiate de la PPI, et 93% pour *ta gueule* alors que ce sont seulement 54% des verbes introduisant *tu te rends compte* et 37% des verbes introduisant *pas vrai*. L'analyse sera notamment poursuivie en étudiant si ces verbes diffèrent en fonction de leur position par rapport à la PPI.

Ces premiers résultats seront approfondis en poursuivant le codage des verbes introducteurs de discours et des PPI, afin de pouvoir analyser la façon dont sont combinés ces deux éléments et étudier les différentes formes et fonctions que peuvent avoir les séquences introductrices de discours direct dans le roman français contemporain.

Références bibliographiques

Anscombre, J.-C. (2015). Verbes d'activité de parole, verbes de parole et verbes de dire : des catégories linguistiques ? *Langue française*, 186, 103-122.

Diwersy, S., Gonon, L., Goossens, V., Gymnich, M., Tutin, A. (2020). Direct Speech in French and English Novels. In I. Novakova & D. Siepmann (éds.), *Phraseology and style in subgenres of the novel: a synthesis of corpus and literary perspectives*, 83-113. London, Palgrave Mcmillan.

Durrer, S. (1999). *Le dialogue dans le roman.* Paris, Nathan.

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

Novakova, I., Siepmann, D. (eds) (2019). *Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives*. London, Palgrave Macmillan.

Harras, G., Winkler, E., Erb, S., Proost, K. (2004). *Handbuch deutscher Kommunikationsverben*. Teil I: Wörterbuch. Berlin, Walter de Gruyter.

Peng, Q. Yuhao, Z., Yuhui, Z., Jason, B., Manning, C. D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.* arXiv:2003.07082. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Rosier, L. (2008). *Le discours rapporté en français*. Paris, Ophrys.

Salvan, G. (2005). L'incise de discours rapporté dans le roman français du XVIII^{e} au XX^{e} siècle : contraintes syntaxiques et vocation textuelle. In A. Jaubert (éd.), *Cohésion et cohérence* (1-). ENS Éditions.

Tutin, A., Gharbi, N. (2020). Phrases préfabriquées à fonction expressive dans les dialogues de romans contemporains. *In*L. Fesenmeier et I. Novakova (éd.) *Phraséologie et stylistique de la langue littéraire*. Peter Lang, 17-29. https://hal.science/hal-03250973

Aider les adultes avec un Trouble du Spectre de l'Autisme à développer des compétences interactionnelles en exploitant des corpus oraux et multimodaux : les projets INTERFAC-TSA et ITSA

Virginie André [1], Carole Etienne [2]
[1] Laboratoire ATILF, Université de Lorraine et CNRS
[2] Laboratoire ICAR, ENS Lyon, CNRS et Université Lyon 2
Virginie.Andre@univ-lorraine.fr, Carole.Etienne@ens-lyon.fr

Introduction

Le trouble du spectre de l'autisme (TSA), tel qu'il est défini par le DSM-V-TR (American Psychiatric Association, 2022), fait partie des troubles neurodéveloppementaux et affecte notamment la communication sociale et les comportements. Par conséquent, les adultes avec un TSA peuvent rencontrer des difficultés pour communiquer et participer à une interaction sociale. Certaines personnes ont du mal à initier ou à maintenir une conversation, à comprendre les codes sociaux, à anticiper les attendus, à prendre la parole ou encore à interpréter les différentes modalités de l'interaction, comme les expressions faciales, les intonations ou les gestes (Attwood, 2018 ; Danon-Boileau & Morel, 2007). En outre, il peut arriver que les tours de parole ne soient pas respectés ou que des silences surviennent, faute de savoir comment participer à un échange. L'ironie, le second degré et les implicites posent aussi problème et constituent des sources de malentendus (Happé, 2005). Ces difficultés peuvent entrainer des incompréhensions, de l'anxiété, des troubles psychologiques et affecter la qualité des relations sociales des adultes autistes (Hollocks et al., 2019 ; Hudson et al., 2019 ; Shtayermman, 2007), que ce soit dans la vie quotidienne, au travail ou encore à l'université. Elles ne sont pas toujours liées à un manque de volonté ou d'intérêt pour les autres, mais à une manière différente de percevoir et de traiter l'information sociale, souvent liée à la théorie de l'esprit, la prise en compte du contexte, la réciprocité sociale et/ou émotionnelle et à la compréhension du fonctionnement des interactions (Baron-Cohen, 1995 ; Frith, 2003 ; Tager-Flusberg et al., 2005).

La prise en charge des enfants avec un TSA est structurée autour de plusieurs champs d'intervention et professionnels de santé (psychologues, orthophonistes, ergothérapeutes, éducateurs spécialisés, etc.) mais cesse le plus souvent à partir de 18 ans, et peu d'actions sont alors mises en place pour aider les adultes (Taylor & Seltzer, 2011). Le problème se pose davantage lorsque les adultes sont diagnostiqués tardivement, c'est-à-dire après leur majorité. Des dispositifs ont été développés pour soutenir les habiletés sociales chez les personnes autistes, en réponse aux difficultés qu'elles peuvent rencontrer dans leurs interactions quotidiennes. Ces accompagnements visent généralement à renforcer la compréhension des

codes sociaux, la régulation émotionnelle, la reconnaissance des intentions et des émotions d'autrui ou encore la prise en compte du contexte dans des situations précises. Ce travail sur les habilités sociales se fait généralement en groupe, à l'aide de jeux de rôle ou d'exercices guidés (Jantz, 2011 ; Saley, 2023). Différents outils et supports peuvent également être utilisés (dessins, BD, visuels, applications numériques, jeux vidéo, etc.) pour favoriser le développement de comportements adaptés. Cependant, très peu de dispositifs d'accompagnement s'appuient sur des vidéos d'interactions authentiques (Michallet et al., 2019), la plupart utilisent des vidéos jouées par des acteurs. De même, les praticiens exploitent peu les résultats des recherches en analyse des interactions malgré les rapprochements qui se mettent en place entre les formations en orthophonie ou en psychologie et les départements de sciences sur langage.

Les projets INTERFAC-TSA et ITSA s'inscrivent dans le cadre de la linguistique de corpus et de l'analyse sociolinguistique des interactions (André, 2025 ; Kerbrat-Orecchioni, 2005 ; Traverso, 2016). Ils élaborent des ressources numériques multimodales, à partir de corpus d'interactions, qui documentent et explicitent le déroulement et les spécificités des interactions sociales identifiées comme problématiques. Dans cette communication, nous présenterons notre approche méthodologique ainsi que certaines des ressources.

Corpus et méthodologie

Corpus

Dans le cadre de ces projets, les ressources visant à aider les adultes autistes à développer des compétences en interaction sont élaborées à partir de corpus oraux et multimodaux d'interactions recueillies de façon écologique. Ainsi, ces ressources exploitent les corpus CLAPI (http://clapi.icar.cnrs.fr/), INTERFARE (https://icar.cnrs.fr/interfare/), TCOF (https://tcof.atilf.fr/) et FLEURON (https://fleuron.atilf.fr/). Le corpus CLAPI (corpus de langue parlée en interaction) est une banque de 67h de données multimédia, majoritairement vidéos, enregistrées en situation réelle, sans intervention ni consigne du chercheur, dans des contextes variés : interactions professionnelles, institutionnelles ou privées, commerciales, didactiques, juridiques ou encore médicales. La plateforme INTERFARE décrit les étapes clés d'une réunion de travail dans une variété de contextes professionnels (200 extraits) et propose une centaine d'exercices multimédias afin de s'entrainer à reconnaitre les procédés et à comprendre les objectifs des participants. Le corpus TCOF (traitement de corpus en français) est composé de 55h d'enregistrements audios de différents genres de discours (conversations, entretiens, réunions, etc.). Le corpus du dispositif d'apprentissage du français FLEURON (français langue étrangère universitaire : ressources et outils numériques) est constitué de 20h de vidéos de situations d'interactions représentatives de la vie d'un étudiant sur un campus (à la scolarité, à la bibliothèque, au restaurant universitaire, etc.) et dans la vie quotidienne (dans les commerces, entre amis, dans des lieux de cultures, etc.).

En dehors des entretiens, les interactions utilisées sont toutes des interactions réelles qui auraient eu lieu même si elles n'avaient pas été enregistrées.

Méthodologie

Les travaux en analyse des interactions ont permis de décrire le français parlé dans des situations privées ou professionnelles variées, d'après les pratiques langagières mobilisées par les locuteurs (Kerbrat-Orecchioni, 2001). Ces analyses ont mis en évidence un certain nombre de routines pour réaliser des fonctions ordinaires du langage, mais elles ont également montré et explicité la variété des procédés verbaux et multimodaux que les participants peuvent mettre en œuvre pour mener à bien leurs échanges en s'adaptant continuellement aux autres

participants et à la situation de communication (anonyme & anonyme, 2022 ; Mondada, 2017 ; Quignard et al., 2016 ; Traverso, 2016).

Afin de construire des ressources adaptées aux besoins des personnes présentant un TSA, nous avons diffusé un questionnaire pour recenser leurs difficultés et les sujets qui demanderaient à être approfondis avec des questions ouvertes pour bien cerner la nature de leurs troubles.

Dans le cadre de nos projets, nous décrivons plusieurs éléments spécifiques des interactions ainsi que leurs valeurs et leurs effets pragmatiques afin d'apporter un éclairage sur leur fonctionnement. Les éléments étudiés sont notamment : la multimodalité (intonations, expressions faciales, regards, gestes, postures), les prises de tour de parole (interruption, chevauchement, pauses), les marqueurs de discours (accord, désaccord, implication de son interlocuteur, introducteur, organisationnel), les implicites (questions, suggestions, requêtes), l'humour ou encore les répétitions.

Résultats

Les projets complémentaires INTERFAC-TSA et ITSA ont permis d'élaborer des ressources différentes afin de développer les compétences en interaction de plusieurs publics.

Le projet INTERFAC-TSA, dédié aux étudiants et aux personnels des universités, propose :

- Des rubriques au sein desquelles des activités langagières sont décrites et explicitées précisément. Par exemple, dans la rubrique « Interagir au téléphone », les activités « Prendre un rendez-vous », « Inviter » et « Faire une réclamation » sont explicitées et illustrées en contexte.
- Des fiches explicatives des activités langagières. Par exemple, la fiche « Demander des renseignements dans un commerce » reprend de manière synthétique les étapes et les explications puis formule des recommandations pour éviter les situations bloquantes.
- Un module de sensibilisation destiné aux interlocuteurs des personnes autistes (enseignants, étudiants, personnels, tuteurs de stage, collègues, etc.) afin de pointer leurs difficultés potentielles et de suggérer des pratiques langagières plus appropriées.

Le projet ITSA met à la disposition des professionnels de santé et des adultes autistes :

- Des fiches descriptives pour les professionnels de santé et les accompagnants qui ont besoin à la fois de données authentiques et d'explications sur le déroulement des interactions (séquences et activités langagières) et les spécificités du français parlé en interaction. L'objectif est de permettre à ces professionnels de co-analyser une vidéo d'interaction complète (de l'ouverture à la clôture) avec les accompagné·es, selon les principes de la *data-session* (Stevanovic & Weiste, 2017).
- Des vidéos interactives pour les adultes autistes auxquelles sont ajoutées des questions, des explications, des activités d'observation et d'analyse afin de développer des capacités métacognitives et métalangagières transférables dans des situations réelles futures. Ces vidéos sont consultables en autonomie et exposent les utilisateurs à une variété de situations de communication.

Les ressources élaborées dans le cadre de ces projets viennent combler un manque important et souvent relevé par les professionnels de santé, celui de l'exploitation de données authentiques, non jouées et non scriptées, permettant d'exposer les personnes avec un TSA à la langue réelle.

Les premières expérimentations des ressources élaborées dans INTERFAC-TSA et ITSA sont prometteuses. Elles pointent leur utilité, leur originalité et leur pertinence.

Références bibliographiques

American Psychiatric Association (Éd.). (2022). *Diagnostic and statistical manual of mental disorders : DSM-5-TR* (Fifth edition, text revision). American Psychiatric Association Publishing.

André, V. (2025). Approche matérialiste et sociolinguistique des interactions verbales. *La Pensée*, *421*(1), 16-26.

anonyme, x., & anonyme, x. (2022). Décrire la syntaxe dans les interactions du quotidien : Objectifs et implications dans l'enseignement et l'apprentissage du français langue étrangère : *Travaux de linguistique*, *n° 84-85*(1), 191-210. https://doi.org/10.3917/tl.084.0191

Attwood, T. (avec Schovanec, J.). (2018). *Le syndrome d'Asperger : Guide complet* (4e éd). De Boeck.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. MIT press.

Danon-Boileau, L., & Morel, M.-A. (2007). Approche linguistique du discours autistique : Quelques remarques. In *Langage, voix et parole dans l'autisme* (p. 335-340). Presses Universitaires de France. https://doi.org/10.3917/puf.touat.2007.01.0335

Frith, U. (2003). *Autism: Explaining the enigma*. Blackwell publishing. https://psycnet.apa.org/record/2003-00578-000

Happé, F. (2005). *Autism: An introduction to psychological theory*. Psychology Press. https://www.taylorfrancis.com/books/mono/10.4324/9780203014226/autism-francesca-happ%C3%A9

Hollocks, M. J., Lerh, J. W., Magiati, I., Meiser-Stedman, R., & Brugha, T. S. (2019). Anxiety and depression in adults with autism spectrum disorder: A systematic review and meta-analysis. *Psychological Medicine*, *49*(4), 559-572. https://doi.org/10.1017/S0033291718002283

Hudson, C. C., Hall, L., & Harkness, K. L. (2019). Prevalence of Depressive Disorders in Individuals with Autism Spectrum Disorder : A Meta-Analysis. *Journal of Abnormal Child Psychology*, *47*(1), 165-175. https://doi.org/10.1007/s10802-018-0402-1

Jantz, K. M. (2011). Support Groups for Adults With Asperger Syndrome. *Focus on Autism and Other Developmental Disabilities*, *26*(2), 119-128. https://doi.org/10.1177/1088357611406903

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

Kerbrat-Orecchioni, C. (2005). *Le discours en interaction*. Armand Colin.

Michallet, B., Taylor, J., Dumont, C., McIntyre, J., & Couture, M. (2019). La communication et les relations interpersonnelles des adultes présentant un trouble du spectre de l'autisme : Une revue systématique des programmes d'intervention. *Revue de psychoéducation*, *48*(1), 117-146. https://doi.org/10.7202/1060009ar

Mondada, L. (2017). Le défi de la multimodalité en interaction. *Revue française de linguistique appliquée*, *Vol. XXII*(2), 71-87. https://doi.org/10.3917/rfla.222.0071

Quignard, M., Ursi, B., Rossi-Gensane, N., André, V., Baldauf-Quilliatre, H., Etienne, C., Plantin, C., & Traverso, V. (2016). Une méthode instrumentée pour l'analyse multidimensionnelle des tonalités émotionnelles dans l'interaction. *SHS Web of Conferences*, *27*, 15004. https://doi.org/10.1051/shsconf/20162715004

Saley, E. (2023). *Entraînement aux habilités sociales. Pour adolescents et adultes autistes*. Autisme diffusion.

Shtayermman, O. (2007). Peer Victimization in Adolescents and Young Adults Diagnosed with Asperger's Syndrome: A Link to Depressive Symptomatology, Anxiety Symptomatology and Suicidal Ideation. *Issues in Comprehensive Pediatric Nursing*, *30*(3), 87-107. https://doi.org/10.1080/01460860701525089

Stevanovic, M., & Weiste, E. (2017). Conversation-analytic data session as a pedagogical institution. *Learning, Culture and Social Interaction*, *15*, 1-17. https://doi.org/10.1016/j.lcsi.2017.06.001

Tager-Flusberg, H., Paul, R., & Lord, C. (2005). Language and communication in autism. In F. R. Volkmar, R. Paul, A. Klin, F. R. Volkmar, & D. Cohen (Éds.), *Handbook of Autism and Pervasive Developmental Disorders* (Third Edition, Vol. 1, p. 335-364). Wiley. https://doi.org/10.1002/9780470939345

Taylor, J. L., & Seltzer, M. M. (2011). Employment and Post-Secondary Educational Activities for Young Adults with Autism Spectrum Disorders During the Transition to Adulthood. *Journal of Autism and Developmental Disorders*, *41*(5), 566-574. https://doi.org/10.1007/s10803-010-1070-3

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

Aménager une interaction institutionnelle formatée : analyse de procès verbaux de plainte pour violence conjugale rédigés avec un « masque de plainte »

Julie Lefebvre
Laboratoire MoDyCo, Université Paris Nanterre
jlefebvre@parisnanterre.fr

La présente communication prend pour objet les interactions entre un policier et une victime telles qu'elles sont représentées (Authier-Revuz et Lefebvre 2015), à l'écrit, dans des procès verbaux de plainte pour violence conjugale produits par la Brigade Locale de Protection de la Famille (BLPF) d'un commissariat de police parisien. Il s'agit plus précisément de mener une étude exploratoire de la représentation écrite d'une interaction institutionnelle orale qui apparaît comme extrêmement formatée. En effet, depuis le Grenelle de lutte contre les violences conjugales, et suite aux préconisations du Centre Hubertine Auclerc (2019), les policiers sont tenus d'utiliser un « masque de plainte » pour les auditions relatives aux violences conjugales, ce mode rédactionnel garantissant une plus grande homogénéité et une plus grande exhaustivité des textes, et favorisant leur exploitation au cours de la procédure.

Le masque de plainte consiste en un canevas textuel qui se présente sous la forme d'un fichier texte généré par le Logiciel de Rédaction de Procédure de la Police Nationale. Il est pré structuré en différentes sections ; dans l'ordre : « Identité et situation de la victime », « Identité et situation du mis en cause », « Sur les faits », « Sur les traces, indices et témoins », « Sur la situation antérieure aux faits », « Dépôt de plainte » et « Contacts partenaires – Informations – Droits ». Hormis les deux dernières, chacune de ces sections contient une liste de questions pré rédigées, qui appellent une réponse de la victime sous forme de ligne restant à remplir et intitulée « Réponse ». La section « Sur les faits », centrale, est quant à elle initiée par la question « Pouvez-vous nous parler des faits que vous souhaitez dénoncer ? » et la réponse avec laquelle elle fait paire est accompagnée du commentaire « Récit libre. Demander ensuite une description détaillée du contenu », qui fait de cette section la seule où le policier est enjoint à développer librement ses questions.

Du point de vue de la « mise en texte » (Rock 2017), la rédaction du procès-verbal s'effectuant en même temps que l'audition et donc simultanément ou quasi simultanément à l'interaction orale entre le policier et la victime, le masque de plainte sert en général de support à cette interaction orale qui va être représentée à l'écrit. Cependant, l'habileté professionnelle du gardien de la paix, liée à son ancienneté comme à ses connaissances en matière de prise en charge des violences conjugales, lui permet d'user du masque de plainte avec plus ou moins de souplesse, autrement dit l'autorise à intervenir, à l'oral, sur l'organisation et le contenu de

l'interaction tels qu'ils sont définis dans le formulaire imposé, écart dont il pourra ensuite rendre compte dans le procès verbal écrit. Ce sont ces aménagements du cadre interactionnel formaté par le masque de plainte, observables dans les procès verbaux de plainte, dont nous souhaitons amorcer l'analyse dans la présente communication. Plus précisément, pour faire émerger des types d'aménagement du cadre interactionnel, nous comparerons de ce point de vue les procès verbaux rédigés par des gardiens de la paix expérimentés, avec des procès verbaux rédigés par des gardiens de la paix nouvellement arrivés dans un service spécialisé en matière de prise en charge des violences conjugales. Il s'agira de repérer, d'une part, les « lieux » (Lefebvre 2022), en l'occurrence les sections du procès verbal dans lesquelles ces sorties hors du masque se produisent. D'autre part, on questionnera la pertinence de l'utilisation de la catégorie du « motif » (Née et. al. 2017 : 118) pour interroger la forme textuelle qu'est le masque de plainte.

Le corpus sera constitué de 30 procès verbaux de plainte pour violence conjugale, constitués de 5 à 11 pages, anonymisés et collectés en mai-juillet 2024 et en juin 2025 au sein de la Brigade Locale de Protection de la Famille d'un commissariat parisien, après autorisation des victimes, avec accord du Parquet de Paris et dans le cadre d'une convention établie avec la Préfecture de Police de Paris. Le corpus sera partitionné en fonction du degré d'ancienneté des gardiens de la paix scripteurs dans cette BLPF : gardiens de la paix expérimentés en poste à la BLPF depuis au moins un an et pour certains depuis plusieurs années / gardiens de la paix arrivés à la BLPF depuis moins d'un an, depuis quelques semaines pour certains.

Les textes du corpus étant imprimés, ils nécessitent une chaîne de numérisation complète : scan, océrisation, correction, conversion et encodage xml. L'annotation prend en compte la structuration du texte en sections ainsi qu'en paires question-réponse. Nous distinguerons les paires question-réponse inscrites originellement dans le masque de plainte des paires question-réponse ajoutées au texte du masque ainsi que de tout élément inséré (didascalies par exemple) ou modifié (pronoms et genre grammatical par exemple) par le policier rédacteur dans le texte du masque du procès-verbal. Sur cette base, nous prévoyons d'effectuer des analyses exploratoires avec le logiciel TXM.

Trois types de résultats sont attendus :

- Au niveau des sections constituant le texte du procès verbal : en dehors de la section « Sur les faits », seront mises en évidence les sections et les questions qui sont le plus fréquemment les lieux d'intervention de la part des gardiens de la paix.
- Au niveau de l'intervention sur le texte, nous nous intéresserons à la forme linguistique des aménagements effectués en distinguant modification de la formulation d'une question et insertion d'une question inédite. Dans ce dernier cas, nous porterons plus particulièrement notre attention sur la nature de l'enchaînement entre l'interaction pré formatée et l'interaction libre entre le policier et la victime.
- Enfin, on mettra en relation ces analyses avec le degré d'expérience du policier scripteur en matière de rédaction de procès verbaux de plainte pour violence conjugale. Par là, il s'agira de mettre au jour différents aménagements du texte formaté par le masque de plainte et d'évaluer leur pertinence et leur nécessité, d'une part eu égard au déroulement de l'interaction entre le policier et la victime, d'autre part relativement à la suite de la procédure.

Références bibliographiques

Authier-Revuz, J. et Lefebvre, J. (2015). L'entretien de presse : un genre discursif de représentation de discours autre. *Revista Investigacoes- Linguística e Teoria Literária*, Universidade Federal de Pernambuco, 28, http://www.repositorios.ufpe.br/revistas/index.php/INV/article/view/1840/1455.

Centre Hubertine Auclerc (2019). *Diagnostic collaboratif sur l'accueil des femmes victimes de violences conjugales et/ou sexuelles et l'évaluation du danger dans trois commissariats de Paris et de la petite couronne*.

Lefebvre, J. (2022). *La Note de bas de page dans les imprimés contemporains*. Limoges: Lambert-Lucas.

Née, E., Leblanc J.-M. et Fleury S. (2017). Compter dans les textes, quelles unités ? Dans Née, E. (éd.), *Méthodes et outils informatiques pour l'analyse des discours*. Rennes : Presses universitaires de Rennes, 103-121.

Rock, R. (2017). Recruiting Frontstage Entextualization: drafting, artifactuality and writtenness as resources in police - witness interviews. *Text & Talk* 37: 4, 437-460.

Analyse pragmatique et discursive de la réponse évasive dans un corpus d'interviews politiques

Antonia Sánchez Villanueva
Université d'Almeria (Espagne)
sva933@ual.es

Introduction

L'idée selon laquelle les politiciens recourent fréquemment à des techniques d'évitement afin d'éluder des réponses claires et sans ambiguïté à des questions délicates est profondément ancrée dans les croyances sociales (Harris, 1991 : 76), au point d'être devenue un lieu commun. La recherche académique dans le domaine de la linguistique, plus précisément dans le champ de l'analyse du discours, s'est progressivement intéressée à cette pratique discursive depuis la fin des années 1980, dans le but de décrire ses mécanismes, de quantifier sa récurrence et d'approfondir les causes ayant contribué à en faire une étiquette négative associée à l'image des personnalités politiques. En effet, d'un point de vue pragmatique en termes de coopération verbale, l'évitement de la réponse au cours d'une interaction constitue d'emblée une violation manifeste des maximes de la coopération verbale telles que formulées par Grice (1979 : 61), notamment celle de la relation, qui suppose un principe de pertinence dans les énoncés du locuteur.

Peut-être, comme le suggère Charaudeau (2005), l'homme politique se trouve-t-il irrémédiablement contraint, par la force des circonstances, à ne pas dire tout ce qu'il sait ou pense à chaque instant, afin d'éviter de nuire à son image ou de compromettre sa propre action future. Ainsi, il ne s'agirait pas tant de présupposer que le politicien dit la vérité, mais plutôt qu'il **donne l'impression de dire la vérité**. Dans les premiers travaux de Bavelas et al. (1988, 1990) sur le langage équivoque, le cadre situationnel est déjà clairement identifié comme le principal facteur explicatif d'une pratique aussi répandue, supplantant les interprétations qui privilégiaient les aspects psychologiques des individus.

Dans cette perspective, un genre de discours tel que l'entretien politique télévisé en direct crée des conditions qui mettent en jeu l'image de la personnalité politique. Il s'agit d'un dispositif de communication dans lequel le contrôle de l'interaction (prises de parole, enchaînement thématique) incombe aux intervieweurs, tandis que les interviewés (les acteurs politiques, en l'occurrence) doivent se limiter à répondre. Cependant, le discours produit par un politique dans ce contexte est formulé en réponse aux questions de l'intervieweur, tout en étant orienté vers l'audience absente. En termes pragmatiques, le statut participatif de cette audience est celui d'un destinataire collectif, hétérogène et secondaire ou latéral (Kerbrat-Orecchioni, 2016 : 56), pour lequel la valeur des actes de parole produits au cours de l'échange peut différer de celle attribuée aux mêmes actes par le destinataire principal ou allocutaire, c'est-à-dire l'(les) intervieweur(s). D'où la multiplicité et la variabilité des portées que peut prendre le discours politique dans ce type de situation de communication, ce qui accroît les risques pour l'image et prédispose à l'usage de formules d'évitement.

L'objectif de cette communication est d'analyser, dans une perspective pragmatique, les stratégies d'évitement employées dans ce contexte de communication par les présidents de la V[e] République dans le cadre des entretiens institutionnels télévisés du 14 juillet. Final del formulario

Conçus comme un rendez-vous régulier où le président fait face aux questions de plusieurs journalistes, ces entretiens s'inscrivent dans le genre des interviews informatives ou *news interviews* (Clayman et Heritage, 2002). Du point de vue de leur ancrage social, elles font partie des genres conversationnels «sollicités et non polarisés» (Charaudeau, 2014).

Corpus et méthodologie

Le corpus d'analyse est constitué par les 35 entretiens télévisés accordés depuis 1978 par les présidents de la V[e] République à l'occasion du 14 juillet. Tous sont disponibles en accès libre, dans leur version retranscrite ou en format vidéo, sur le site web officiel www.vie-publique.fr. Il s'agit d'un corpus que nous avons nous-mêmes délimité à partir d'un ensemble homogène d'interviews d'une durée moyenne de 55 minutes, totalisant près de 250 000 mots, ce qui le définit comme un corpus de taille moyenne (Moirand 2018). Ce recueil est par ailleurs non clos, susceptible de s'accroître à l'avenir et dont le traitement peut être abordé aussi bien en synchronie qu'en diachronie. Par conséquent, en tant que corpus bien délimité, les interviews présidentielles du 14 juillet peuvent être considérées comme «la construction d'un dispositif d'observation propre à révéler, à faire appréhender l'objet de discours qu'elle se donne pour tâche d'interpréter» (Mazière, 2018 : 9)

Méthodologie

Notre méthode d'analyse combine des procédures quantitatives et qualitatives, et repose sur une série de catégories que nous avons nous-mêmes élaborées.

1º) Délimitation des unités conversationnelles

Pour chacune des interviews, un travail d'identification et de segmentation des unités conversationnelles de base a été réalisé, afin de délimiter les paires adjacentes question-réponse au sein de la dynamique interactionnelle.

2º) Délimitation des questions

Pour chaque interview, les interventions constituant un acte de parole interrogatif ont été identifiées, quantifiées et classées.

3º) Délimitation des réponses/non-réponses

À partir des transcriptions, les interventions constituant, dans le cadre d'une paire adjacente, une réponse réactive à la question ont été identifiées et codifiées manuellement à l'aide du logiciel NVivo.

4º) Quantification et classification des réponses évasives

À partir de cette étape, l'analyse des données s'est centrée sur les items identifiés comme [N-R], auxquels nous avons appliqué la classification des réponses évasives : celle proposée par Bull et Mayer (1993), complétée par Bull (2003), puis modifiée et élargie par nos soins à partir de la casuistique observée dans le corpus.

5º) Analyse comparative des données

Les données ont été saisies dans des fichiers Excel en vue de leur traitement statistique.

6º) Analyse des résultats

La dernière étape consiste en une analyse qualitative avec Nvivo des résultats obtenus, accompagnée d'un commentaire des traits pragmatiques et discursifs de fragments sélectionnés du corpus.

Résultats

Les données extraites du corpus révèlent un comportement évasif de la part des présidents français, similaire à celui déjà observé par d'autres études du monde anglophone chez les responsables politiques nord-américains et britanniques. Ainsi, les réponses évasives représentent près de 50 % de leurs interventions. Par ailleurs, les données mettent en évidence une relation de cause à effet étroite entre la contrainte imposée par la question et l'évitement de la réponse : plus le champ discursif est restreint et plus la formulation de la question est polarisée, plus la probabilité d'une réponse évasive augmente.

Références bibliographiques

Bull, P. (2003). *The microanalysis of Political Communication: Claptrap and Ambiguity*. Routledge. https://doi.org/https://doi.org/10.4324/9780203417843

Bull, P. (2008). Slipperiness, Evasion, and Ambiguity: Equivocation and Facework in Noncommittal Political Discourse. *Journal of Language and Social Psychology*, *27*(4), 333–344. https://doi.org/10.1177/0261927X08322475

Bull, P. et Mayer, K. (1993). How not to answer questions in political interviews. *Political Psychology*, 14(4), 651-666. https://doi.org/10.2307/3791379

Charaudeau, P. (2005). *Le discours politique. Les masques du pouvoir*. Vuibert.

Charaudeau, P. (2014). «La situation de communication comme fondatrice d'un genre: la controverse. En M. Monte M. et Ph. Gilles (Éds.), *Genres & Textes: Déterminations, évolutions, confrontations*, (pp 49-57). Presses universitaires de Lyon.

Clayman, S. et Heritage, J. (2002). *The News Interview: journalists and public figures on the air*. Cambridge University Press. https://doi.org/10.1017/CBO9780511613623.

Ekström, M., et Tolson, A. (2017). Political Interviews: Pushing the Boundaries of Neutralism.' In M. Ekström y Firmstone J. (Eds.), *The Mediated Politics of Europe* (pp. 123–149). Springer International Publishing. https://doi.org/10.1007/978-3-319-56629-0_5

Ekström, M., Eriksson, G., Johansson, B., et Wikström, P. (2013). Biased interrogations? *Journalism Studies*, *14*(3), 423–439. https://doi.org/10.1080/1461670X.2012.689488

Ezpeleta Piorno, P. et Gamero Pérez S. (2004). Los géneros técnicos y la investigación basada en corpus: proyecto GENTT. In R. Gaser, C. Guirado y J. Rey, Insights into Scientific and Technical Translation, pp. 147-156. PPU-Universitat Pompeu Fabra.

Grice, H. P. (1979). Logique et conversation. *Communications. La Conversation*, *30*(1), 57–72. https://doi.org/DOI : https://doi.org/10.3406/comm.1979.1446

Harris, S. (1991). Evasive action: how politians respond to questions in political interviews. En P. Scannell (Ed.), *Broadcast Talk* (pp. 76–99). Sage Publications.

Kerbrat-Orecchioni, C. (1998). La notion d'interaction en linguistique: origine, apports, bilan. *Langue Française*, 117, 51–67. https://doi.org/10.3406/lfr.1998.6241

Kerbrat-Orecchioni, C. (2016). *Les actes de langage dans le discours*. Armand Colin.

Leblanc, J.M. et Martinez, W. (2005). « Positionnements énonciatifs dans les vœux présidentiels sous la cinquième République ». *Corpus*, 4. https://doi.org/10.4000/corpus.347

Mazière, F. (2018). *L'analyse du discours. Histoire et pratiques*. *Que sais-je ?* Presses Universitaires de France. https://www.cairn.info/l-analyse-du-discours--9782130813958.htm.

Moirand, S. (2018). L'apport de petits corpus à la compréhension des faits d'actualité. *Corpus* 18.

Sánchez Villanueva, A. (2022). Hacia una teoría pragmática de la respuesta evasiva en el discurso político francés y español. Estudio comparativo-tipológico de entrevistas televisadas (Tesis doctoral). Universidad de Almería. https://repositorio.ual.es/handle/10835/13853

Turbide, O. (2015). La construction d'images publiques dans le discours politique médiatique. *Communiquer*, *14, 5-23*. https://doi.org/10.4000/COMMUNIQUER.1624

Annotation linguistique semi-automatique en contexte réel : développement et test d'un système de transcription phonémique automatique pour les dialogues en thulung

Séverine Guillaume [1], Guillaume Wisniewski [2] et Aimée Lahaussois [3]
[1] LACITO, CNRS, Université Sorbonne Nouvelle, F-94800 Villejuif, France
[2] Université Paris Cité, LLF, CNRS, F-75013 Paris, France
[3] Université Paris Cité, HTL, CNRS, F-75013 Paris, France
severine.guillaume@cnrs.fr, guillaume.wisniewski@u-paris.fr, aimee.lahaussois@cnrs.fr

Contexte et enjeux

La révolution des giga-modèles de langue (LLMs) ouvre de nouvelles perspectives pour la documentation linguistique. Elle permet notamment de concevoir des systèmes capables d'annoter automatiquement des textes dans des langues ou des contextes pour lesquels il n'existe que très peu de données annotées (Anastasopoulos, Cox, Neubig, & Cruz, 2020 ; Partanen, Hämäläinen, & Klooster, 2020). Ces modèles pourraient donc faciliter le travail des linguistes en automatisant, au moins en partie, des tâches d'annotation ou de transcriptions qui, si elles sont essentielles, sont connues pour être chronophages et fastidieuses. Ainsi, dans nos travaux antérieurs (Macaire et al., 2021 ; Guillaume et al., 2022), nous avons montré, dans le contexte de documentation de langues rares et en danger, qu'il était possible de développer des systèmes de transcription phonémique de qualité satisfaisante à partir de seulement quelques heures de données transcrites.

Dans cette étude, nous présentons une nouvelle série d'expériences visant à développer un système de transcription phonémique pour le thulung, une langue tibéto-birmane parlée au Népal, encore peu documentée, en considérant des textes de nature plus diverses que ce qui avait été fait jusqu'à présent. Nous nous appuyons pour cela sur un corpus de 12 h d'enregistrement en thulung collecté et transcrit manuellement par une linguiste experte de cette langue, aidée de ses locuteur·trice·s au cours de 10 campagnes sur le terrain entre 1999 et 2025. Les données sont disponibles dans la collection Pangloss (Adamou, Guillaume, & Michaud, 2025), une archive de langues rares et en cours de documentation, disponible librement en ligne.[13]

Comme dans nos travaux précédents, le système développé repose sur l'affinage *fine-tuning* d'un modèle pré-entrainé multingue de la parole, *XLSR-53*. C'est une approche aujourd'hui classique en TAL qui consiste à adapter ce modèle à une langue ou un domaine spécifique en le ré-entraînant sur un jeu de données ciblé. Cela permet de transférer les connaissances générales apprises sur de nombreuses langues (53 dans le cas du modèle considéré) vers une

13 https://www.pangloss.cnrs.fr

tâche ou une langue spécifique, tout en nécessitant moins de données annotées. De nombreux travaux ont montré que cette approche permet d'obtenir d'excellentes performances de transcription, même pour des langues peu dotées ou des contextes particuliers, sans devoir entraîner un modèle depuis zéro.

Annotation semi-automatique : gain de productivité ou perte de temps ?

Notre démarche se distingue de l'état de l'art sur plusieurs points. Tout d'abord, nous avons utilisé le système de transcription développé dans le cadre de ce travail lors d'une véritable campagne d'annotation, ce qui nous permet d'évaluer son utilité dans un contexte appliqué, en dehors des protocoles expérimentaux classiques généralement utilisés dans les travaux portant sur la documentation linguistique computationnelle. Plus précisément, nous avons mené des expériences dans lesquelles la linguiste interagit directement avec les prédictions du système, au cours de son travail de terrain, que ce soit pour les corriger ou les valider. Cela nous permet de mesurer concrètement l'impact du système sur le temps d'annotation, plutôt que de nous limiter, comme c'est souvent le cas des travaux en TAL, à l'évaluation de sa capacité à reproduire une annotation de référence.

Nos résultats, résumés dans la table 1, sont sans appel : si l'utilisation d'un système de transcription automatique permet un gain de temps pour les enregistrements les plus simples, notamment lorsqu'ils sont accompagnés d'annotations préexistantes du même locuteur ou de la même locutrice, ce gain devient nul, voire négatif, pour les textes plus complexes, en particulier les conversations et la parole spontanée (par opposition aux discussions « guidées »).

Type enregistrement	Enregistrement	Correction	Gain de temps
réaction à un stimulus visuel, 1 locutrice	1 mn 55	9 minutes	grand
réaction à un stimulus visuel, 1 locutrice	4 mn 28	9 minutes	grand
lecture, 1 seule locutrice	7 mn 30	29 minutes	moyen
conversation spontanée, 1 locutrice	6 mn 38	80 minutes	faible
conversation spontanée, 4 locuteur·trice·s	4 mn 17	95 minutes	faible voire inexistant

table 1. : Temps de correction de la transcription automatique pour différents enregistrements.

Notons toutefois que ces conclusions reposent sur le ressenti (nécessairement subjectif) de la linguiste (et de ses locuteur·trice·s) ayant réalisé les annotations finales. Toutefois, comparer précisément le temps requis pour corriger une transcription automatique à celui nécessaire pour effectuer une annotation « à partir de zéro » impliquerait de mobiliser deux linguistes compétent·e·s et disponibles pour effectuer ce travail, transcrire un même texte une deuxième fois étant une tâche naturellement plus simple. Cette situation est quasiment impossible à mettre en place dans la pratique, surtout pour des langues rares et en cours de documentation.

Transcrire automatiquement les conversations

Les résultats que nous venons de présenter illustrent un deuxième point de différence de ce travail avec les travaux de l'état de l'art : nous avons testé le système sur des données particulièrement variées. Contrairement à de nombreux travaux (dont les nôtres !) qui s'appuient sur des corpus enregistrés dans des conditions particulièrement favorables pour la

transcription automatique (notamment des corpus mono-locuteur·trice, textes narratifs ou des dialogues simples, sans chevauchement de parole) nos expériences portent sur des corpus hétérogènes incluant des récits spontanés, des conversations naturelles et des locuteur·trice·s multiples.

Nous avons reporté dans la Table 2 une évaluation des performances des systèmes que nous avons développés à partir de différentes combinaisons des données à notre disposition. Il s'agit, comme nous l'avons mentionné, de la manière habituelle d'évaluer les performances d'un système de transcription phonémique en TAL en reportant le CER *character error rate*, c'est-à-dire, la proportion de caractères qu'un·e linguiste devrait modifier pour obtenir une transcription parfaite. Ainsi, un CER de 16,9 % n'est pas un score très impressionnant : il signifie qu'un·e linguiste devrait corriger (insérer, supprimer ou modifier) une moyenne de 17 caractères pour 100 caractères prédits (presque 1 sur 5) !

Corpus apprentissage	Corpus test	CER
1 h mono-locutrice	même locutrice	12,3 %
2 h mono-locutrice	même locutrice	9,7 %
1 h mono-locutrice	autres locuteur·trice·s (mono.)	33,0 %
2 h mono-locutrice	conversation 2 locuteurs·trices	46,1 %
7 h toutes les sources de données	toutes les sources de données	16,9 %

table 2. : CER obtenu par différents systèmes en fonction du type de données utilisées pour apprendre celui-ci et des données sur lesquelles il est évalué (corpus de test).

Ces résultats montrent que les performances du système de transcription varient fortement selon le type de corpus utilisé. En particulier, elles diminuent nettement lorsque les données sur lesquelles le système a été entraîné sont d'un type différent de celui sur lequel il est testé, par exemple lorsqu'un système est appris sur des données mono-locuteur et testé sur des enregistrements multi-locuteur ou lorsque les systèmes sont appliqués sur des locuteur·trice·s non présent·e·s dans les données d'apprentissage. Ce phénomène est bien connu en apprentissage automatique : pour qu'un modèle fonctionne bien, on suppose généralement que les données d'entraînement et de test sont issues de la même distribution — ce qui est rarement le cas en pratique.

Cette chute de performance est particulièrement marquée dans le cas de la transcription de conversations. Deux facteurs principaux peuvent l'expliquer. D'une part, la tâche est intrinsèquement plus complexe : les conversations contiennent souvent des chevauchements de parole, des phrases incomplètes, des hésitations, des reformulations et des interruptions. Elles présentent également une plus grande variabilité de réalisations acoustiques, tant entre les locuteur·trice·s (variabilité *inter*locuteur·trice·s) qu'au sein même de leur discours (variabilité *intra*locuteur·trice). D'autre part, on dispose de moins de données annotées pour ce type de contenu, ce qui limite intrinsèquement les performances du système.

Pour répondre à ces défis, nous présentons plusieurs approches récentes et en cours visant à améliorer la transcription des conversations. Nous essayons notamment d'utiliser des méthodes de diarisation, c'est-à-dire des techniques qui permettent de découper un enregistrement en segments correspondant aux prises de parole de chaque locuteur·trice. Nous avons récemment, dans (Rosina Fernandez, Guillaume, & Wisniewski, 2024), montré que ces méthodes, même appliquées à des langues peu documentées, pouvaient aider à améliorer la qualité de la transcription (et réduire le temps de correction du ou de la linguiste) en préparant automatiquement les données sous forme de tours de parole.

Remerciements

Ce travail a été financé par Université Paris Cité dans le cadre du projet `Twinkle`.

Références bibliographiques

Adamou, E., Guillaume, S., & Michaud, A. (2025). The Pangloss collection : Opening up research data on endangered and underdocumented languages. Language, 101(1), e38–e59. Consulté sur https://muse.jhu.edu/pub/24/article/954237 (Volume 101, Number 1, March 2025)

Anastasopoulos, A., Cox, C., Neubig, G., & Cruz, H. (2020, décembre). Endangered languages meet Modern NLP. In L. Specia & D. Beck (Eds.), *Proceedings of the 28th international conference on computational linguistics : Tutorial abstracts* (pp. 39–45). Barcelona, Spain (Online) : International Committee for Computational Linguistics. Consulté sur https://aclanthology.org/2020.coling-tutorials.7/ doi : M10.18653/v1/2020.coling-tutorials.7

Guillaume, S., Wisniewski, G., Galliot, B., Nguyễn, M.-C., Fily, M., Jacques, G., & Michaud, A. (2022, septembre). Plugging a neural phoneme recognizer into a simple language model : a workflow for low-resource settings. *In Proceedings of Interspeech 2022* (p. 4905-4909). Incheon, South Korea : International Speech Communication Association. Consulté sur https://shs.hal.science/halshs-03625581 doi : M10.21437/Interspeech.2022-11314

Macaire, C., Wisniewski, G., Guillaume, S., Galliot, B., Jacques, G., Michaud, A., … Fily, M. (2021, décembre). Spécialisation de modèles neuronaux pour la transcription phonémique : premiers pas vers la reconnaissance de mots pour les langues rares. *In GDR Lift 2021.* Grenoble, France. Consulté sur https://shs.hal.science/halshs-03475443

Partanen, N., Hämäläinen, M., & Klooster, T. (2020, octobre). Speech recognition for endangered and extinct Samoyedic languages. In M. L. Nguyen, M. C. Luong, & S. Song (Eds.), *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation* (pp. 523–533). Hanoi, Vietnam : Association for Computational Linguistics. Consulté sur Mhttps://aclanthology.org/2020.paclic-1.60/

Rosina Fernandez, C., Guillaume, S., & Wisniewski, G. (2024). Automatisation de la segmentation pour la linguistique documentaire : une nouvelle évaluation des capacités multilingues des modèles neuronaux pré-entrainés de la parole. *In GDR Lift 2024*. Orléans, France.

Approche contrastive intra-linguale : le corpus micro-comparable monolingue comme outil de recherche et d'enseignement

Lidia Lebas-Fraczak[1]
[1] Laboratoire de Recherche sur le Langage, Université Clermont Auvergne
lidia.lebas@uca.fr

Introduction

En linguistique de corpus, la comparaison constitue un outil central pour étudier les phénomènes linguistiques. Deux grands types de corpus sont traditionnellement mobilisés dans ce cadre : les corpus parallèles, qui alignent des textes traduits segment par segment afin d'analyser les équivalences traductionnelles et les phénomènes d'interférence linguistique (Baker, 1995 ; Johansson, 2007), et les corpus comparables, composés de textes similaires produits indépendamment dans chaque langue (McEnery & Xiao, 2007 ; Laviosa, 2002). Ces approches s'appliquent aussi en contexte monolingue, par exemple pour comparer des textes originaux et leurs versions simplifiées (Cardon & Grabar, 2019).

Nous proposons ici un format inédit : le corpus micro-comparable monolingue, constitué de paires d'énoncés authentiques et fortement similaires sur le plan sémantico-discursif, différant par un élément linguistique ciblé. Ce type de corpus permet d'analyser finement des oppositions fonctionnelles internes à une langue, comme, par exemple, l'antéposition et la postposition de l'adjectif épithète en français (*offrant un magnifique panorama* / *offrant un panorama magnifique*), en s'inscrivant dans la perspective de la linguistique de discours comparative telle que définie par S. Moirand (1992), où l'on confronte des objets discursifs présentant à la fois des invariants et des phénomènes de variance. Ce type de corpus peut également constituer un support pédagogique pour l'enseignement universitaire ou le FLE, notamment dans une approche inductive (Nonnon, 1999) ou de la conceptualisation grammaticale (Portine, 2018).

Corpus et méthodologie

Corpus

Le corpus exploratoire a été constitué dans le cadre d'un cours de master SDL par six étudiants travaillant en binômes. Trois adjectifs épithètes affectifs (Kerbrat-Orecchioni, 1980) – *merveilleux*, *magnifique* et *splendide* – ont été retenus, car fréquents dans les deux positions. Pour chaque adjectif, 10 occurrences en antéposition et 10 en postposition ont été collectées à l'aide de Sketch Engine sur des sites Internet proposant des textes rédigés dans un français standard, grammaticalement correct et idiomatique, dans des contextes éditoriaux variés (médias, guides touristiques, forums). Ainsi, le corpus contient 60 énoncés contextualisés organisés en 30 paires présentant un haut degré de similarité contextuelle et discursive, garantissant la comparabilité intra-linguale, comme celle citée ci-dessous.

1) Jour 4 Salvador de Bahia Journée libre en chambre et petit-déjeuner. Suggestion : Départ en bateau à travers la Baie de Todos os Santos jusqu'aux l'îles dos Frades et d'Itarpica. Journée de détente et de baignade de plage en plage offrant un **magnifique panorama** sur Salvador et ses alentours. https://ann.fr/circuit-bresil-essentiels-en-aquarelle-en-onze-jours/	2) Le saviez-vous ? Port Racine, qui tire son nom d'un célèbre corsaire, est le plus petit port de France ! Petit, certes, mais offrant un **panorama magnifique** entre collines verdoyantes et eaux bleu turquoise. https://www.sandaya.fr/villes/la-hague

table 1. : Extrait du corpus exploratoire

Méthodologie

Selon H. Nølke, parmi les facteurs qui influencent la position de l'adjectif, la focalisation est « un facteur très fort » (1996 : 45), si l'on admet « qu'une focalisation a lieu à l'intérieur de chaque syntagme » (*ibid.* : 47). L'auteur propose les règles suivantes (*ibid.* : 48) :

- Un adjectif antéposé ne constitue jamais à lui seul le foyer simple : ou bien il est focalisé avec son substantif ou bien il se trouve en dehors du foyer.
- Un adjectif postposé est toujours focalisé : ou bien il forme le foyer avec son substantif, ou bien il est seul à être focalisé. »

Pour tester ces règles, notre protocole d'analyse applique deux critères à chaque occurrence de syntagme nominal au sein du corpus :

1) Suppression de l'adjectif : (1) si la suppression est grammaticalement possible et n'altère pas la valeur informative ni l'effet pragmatique ou expressif de l'énoncé, l'adjectif est interprété comme non focalisé ; (2) si la suppression est impossible, ou modifie substantiellement le contenu informatif ou l'orientation discursive (par exemple, en supprimant un effet de contraste), l'adjectif est considéré comme focalisé (seul ou avec le substantif).

Ce critère s'appuie sur l'idée, formulée notamment par K. Lambrecht (1994), qu'un élément focalisé ne peut être omis sans priver l'énoncé d'une partie de sa contribution informative, comprise ici au sens large, incluant la dimension pragmatique et l'engagement énonciatif.

2) Présupposition du référent : si le référent du syntagme est présupposé (connu ou accessible dans le contexte discursif), le substantif est interprété comme non focalisé.

Ce critère s'appuie sur l'opposition classique entre information présupposée (donnée) et information focale (nouvelle) (Lambrecht, 1994 ; Krifka, 2007).

L'application de ces deux critères permet de tester empiriquement les prédictions de la théorie de H. Nølke. En particulier, le premier critère suffit à fournir un cas potentiellement falsifiant : la présence d'un adjectif postposé dont la suppression est possible contredit directement la règle selon laquelle tout adjectif postposé appartient au foyer.

Le second critère complète cette analyse en précisant les cas particuliers : (1) en antéposition, la présupposition du référent d'un syntagme dont l'adjectif est supprimable suggère que le nom n'est pas focalisé, ce qui contredit la théorie, sauf à admettre la possibilité de (re)focaliser un référent présupposé ; (2) en postposition, la présupposition permet de distinguer si l'adjectif est focalisé seul (référent présupposé) ou avec le nom (référent non présupposé).

En appliquant ces critères à la paire d'énoncés citée plus haut, on observe que le référent *panorama* n'est pas présupposé dans les deux cas et que l'adjectif est, en principe, supprimable. Toutefois, dans l'exemple 2, la suppression de l'adjectif postposé altère l'effet de contraste et la mise en valeur du paysage ; nous le considérons donc comme non supprimable, et donc comme focalisé. Etant donné que le référent n'est pas présupposé, l'adjectif est interprété comme focalisé avec le substantif. Dans l'exemple 1, en revanche, le substantif est focalisé seul. Pour cette paire d'énoncés, les prédications de la théorie de H. Nølke sont confirmées.

Résultats

1) Antéposition

Dans 66 % des cas, l'adjectif antéposé est supprimable, ce qui indique qu'il se situe hors du foyer, conformément à la théorie de H. Nølke, selon laquelle l'antéposition peut marquer la défocalisation. Dans les autres cas, l'adjectif est interprété comme focalisé conjointement avec le substantif. Toutefois, trois de ces syntagmes présentent un substantif présupposé, donc non focalisé. Ces occurrences invitent à admettre la possibilité d'une (re)focalisation d'un référent présupposé, afin d'intégrer ces cas au modèle théorique.

2) Postposition

Les résultats apparaissent plus hétérogènes. Le taux élevé de substantifs présupposés, donc non focalisés (40 %), semble confirmer que la postposition favorise la focalisation de l'adjectif. Toutefois, seuls 57 % des adjectifs se révèlent non supprimables, donc focalisés. Les 43 % restants correspondent à des adjectifs supprimables, donc non focalisés, ce qui constitue une contre-évidence nette à la règle selon laquelle tout adjectif postposé serait nécessairement focalisé. Parmi ces syntagmes, deux contiennent un nom présupposé ; l'ensemble du syntagme échappe alors à la focalisation, ce qui soulève également une difficulté théorique.

Ainsi, si l'antéposition se comporte globalement comme prévu, la postposition n'assure pas systématiquement la présence de l'adjectif dans le foyer : le fait que 43 % des adjectifs soient supprimables invite à reconsidérer la valeur systématique que H. Nølke attribue à la postposition. Ces résultats suggèrent que d'autres facteurs influencent le choix de la postposition pour des adjectifs pouvant occuper les deux positions. L'un pourrait être d'ordre diachronique : la linguistique historique montre que, si l'antéposition était majoritaire en ancien français et moyen français, la postposition domine aujourd'hui et tend à constituer la position non marquée (Wilmet, 1986 ; Roadman, 2025). Cette évolution pourrait contribuer à affaiblir la valeur focalisatrice que H. Nølke associe à la postposition. Une reformulation en termes de tendance paraît alors pertinente : la postposition reste une ressource privilégiée pour marquer la focalisation de l'adjectif, mais elle ne l'implique pas nécessairement.

Cette étude montre que le corpus micro-comparable constitue un outil particulièrement pertinent pour tester empiriquement des hypothèses théoriques fines, en offrant un terrain d'observation à la fois contrôlé et ancré dans l'usage authentique. Au-delà de la question de la position de l'adjectif, ce type de corpus ouvre des perspectives pour l'analyse d'autres phénomènes linguistiques, et se présente comme une ressource souple et réplicable, au service de la recherche comme de la pédagogie.

Ce format apparaît comme un support prometteur pour une approche inductive de l'enseignement, tant en linguistique universitaire qu'en FLE. Comparer des paires d'énoncés authentiques et contrastés sur un point précis incite étudiants et apprenants à observer les régularités et les variations en contexte, à formuler des hypothèses sur le fonctionnement de la

langue et à les tester faces aux théories existantes. En FLE, il contribue à dépasser un enseignement prescriptif des « règles » de position de l'adjectif pour donner à voir les usages réels, renforçant ainsi la compétence discursive ainsi que la sensibilité pragmatique et stylistique.

Références bibliographiques

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2), 223-243.

Cardon, R., & Grabar, N. (2019). Construction d'un corpus parallèle à partir de corpus comparables pour la simplification de textes médicaux en français. *Traitement Automatique des Langues*, 61(2), 15-39. https://aclanthology.org/2020.tal-2.2.pdf.

Johansson, S. (2007). *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Kerbrat-Orecchioni, C. (1980), *L'énonciation. De la subjectivité dans le langage*. Paris : Armand Colin.

Krifka, M. (2007), Basic notions of information structure. *Interdisciplinary studies of information structure*, 6, 13-56.

Lambrecht, K. (1994). *Information structure and sentence form. Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Laviosa, S. (2002). *Corpus-based translation studies: Theory, findings, applications*. Amsterdam: Rodopi.

McEnery, T. & Xiao, R. (2007). Parallel and comparable corpora: What is happening? In G. Anderman & M. Rogers (Eds.), *Incorporating corpora: The linguist and the translator* (18-31) Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781853599873-005

Moirand, S. (1992). Des choix méthodologiques pour une linguistique de discours comparative. *Langages*, 105, 28-41.

Nølke, H. (1996). Où placer l'adjectif épithète ? Focalisation et modularité. *Langue française*, 111, 38-58.

Nonnon, E. (1999). Tout un nuage de philosophie condensé dans une goutte de grammaire : interactions verbales et élaboration de règles dans la mise en œuvre d'une « démarche inductive » en grammaire. *Pratiques : linguistique, littérature, didactique*, 103-104, 116-148.

Portine, H. (2018). La conceptualisation grammaticale : entre grammaire artificielle et grammaire mentale. *Recherches en didactique des langues et des cultures*, 15-1. http://journals.openedition.org/rdlc/2662 ; DOI : https://doi.org/10.4000/rdlc.2662.

Roadman, A. (2025). Adjective positioning in French: Diachronic change and language contact. *Isogloss*. *Open Journal of Romance Linguistics*, 11(5)/9, 1-22.

Wilmet, M. (1986). *La détermination nominale*. Paris : Presses Universitaires de France

Approche statistique de la distribution des valeurs du marqueur modal *pouvoir* dans des corpus oraux et écrits

Anna Colli [1], Delphine Battistelli [1]
[1] Laboratoire MoDyCo (UMR 7114 Université Paris Nanterre-CNRS)
acolli@parisnanterre.fr, dbattist@parisnanterre.fr

Introduction

Nous nous intéressons ici à l'analyse de la distribution des valeurs sémantiques du verbe *pouvoir* dans des corpus oraux et écrits. Ce travail prend place dans un cadre plus général qui relève de l'analyse du "profil modal" d'un corpus. Nous définissons le profil modal d'un corpus de textes relativement aux proportions de catégories modales qui y sont exprimées, au TOP20 des marqueurs modaux utilisés et à la proportion des marqueurs polysémiques employés. Son calcul est réalisé à partir du repérage automatique des marqueurs modaux lexicaux réalisé par un outil décrit dans (Battistelli et Étienne, 2024). Dans le cadre de cette approche quantitative de la modalité lexicale, nous avons montré dans (Colli et Battistelli, 2025) l'attention particulière que nous souhaitions apporter au verbe *pouvoir* pour deux raisons au moins : d'une part, il s'agit d'un des deux marqueurs modaux (avec bien) les plus fréquents en commun entre tous les corpus que nous avons pu observer ; d'autre part, du fait qu'il est un marqueur polysémique, il joue un rôle clé dans le calcul du profil modal d'un corpus. Les analyses présentées dans (Colli et al., 2025) ont permis de mettre en lumière une variation significative de l'usage de la modalité dans un corpus d'entretiens sur un événement traumatique par rapport à un autre sur une thématique neutre. Dans la continuité de ces travaux, nous nous intéressons ici à la distribution des différentes valeurs sémantiques du verbe pouvoir et formulons des hypothèses relatives aux corpus dans lesquels sont observées ses occurrences, hypothèses que nous cherchons ensuite à valider à l'aide de tests statistiques. Nous présentons tout d'abord notre cadre théorique pour l'analyse de pouvoir (section 1), puis les corpus et notre méthodologie d'analyse (section 2). Les résultats sont présentés en section 3 et nous terminons par un bilan en section 4.

Cadre théorique d'analyse de la sémantique du marqueur modal *pouvoir*

Pouvoir est un marqueur polysémique car il peut avoir, selon le contexte, des valeurs sémantiques différentes. Nous avons choisi de nous appuyer sur l'approche de (Gosselin, 2010) qui retient 4 valeurs possibles pour ce marqueur. Prenons les exemples (1) à (4). Le verbe pouvoir, dans (1), indique une éventualité, dans (2), une possibilité matérielle donnée au sujet par les circonstances ou une capacité du sujet même, dans (3), la contingence d'un état (appelée sporadicité) et, dans (4), une permission accordée au sujet.

1. Cinq mille euros ça pourrait être pas mal. [ESLO - ENT_1228]

2. Il y a des choses que vous ne pouvez pas faire uniquement avec du verre. [ESLO - ENT_002]

3. On bouche un trou, on fait un nœud avec son mouchoir, qu'on peut transformer en bandeau, en brassard, ou encore en signe d'adieu. [TK2_PEDIA]

4. Être féministe c'est donc être conscient qu'une femme peut refuser un rapport sexuel. [TWEETS]

Plusieurs travaux portant sur le français se sont intéressés à la catégorisation des valeurs sémantiques de ce verbe (e.g. Barbet, 2012 ; Vetters, 2012), mais, comme l'indique (Pamuksaç, 2023), rares sont les travaux qui s'intéressent à une analyse statistique de la distribution de ces valeurs dans les corpus. Pour notre part, nous nous appuyons sur l'analyse de (Gosselin, 2010) et, après une étape de désambiguïsation automatique du verbe *pouvoir*, nous menons une analyse statistique de la distribution de ses valeurs sémantiques dans des corpus de nature différente.

Corpus et méthodologie d'analyse

Nous avons retenu deux corpus oraux et deux corpus écrits.

- Corpus oraux : le premier corpus, nommé ici 13_11, est un corpus d'entretiens retranscrits (environ 500.000 tokens), issus de l'Étude 1000 du Programme 13-novembre[1] autour des attentats du 13 novembre 2015 en Île de France. Le deuxième corpus, nommé ici ESLO_ENT, est extrait du corpus ESLO[2]. Il contient 207 entretiens semi-dirigés menés avec les habitants de la ville d'Orléans sur leur quotidien (environ 500.000 tokens).
- Corpus écrits : le premier corpus, nommé ici TWEETS, est un corpus composé de 11.835tweets sexistes ou neutres (environ 400.000 tokens) - extrait de (Chiril et al., 2020). Le deuxième corpus, nommé ici T2K_PEDIA est un corpus de 502 textes encyclopédiques (environ 400.000tokens) - extrait de (Battistelli et al., 2022).

Nous proposons d'explorer deux hypothèses principales :

- Nous faisons l'hypothèse (H1) d'une variation de la distribution des valeurs sémantiques de *pouvoir* dans des corpus du même type avec des contenus différents. En particulier, nous supposons que l'emploi des valeurs diffère entre les deux corpus d'entretiens : ESLO_ENT et 13_11. En effet, dans (Colli et al., 2025), nous avons déjà montré la variation statistiquement significative de l'emploi des marqueurs modaux, dans deux corpus d'entretiens, selon le contenu, traumatique dans le cas de 13_11 ou non traumatique dans le cas de ESLO_ENT.
- Nous faisons par ailleurs l'hypothèse (H2) que la distribution des valeurs sémantiques de *pouvoir* varie significativement entre les corpus oraux et les corpus écrits.

L'approche d'analyse statistique a été menée en deux étapes :

- **Étape 1. Calcul des valeurs sémantiques des occurrences du verbe Pouvoir** : afin de désambiguïser les occurrences du marqueur pouvoir, nous avons appliqué à nos quatre

---

1 https://www.memoire13novembre.fr Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-10-EQPX-0021 Programme 13-Novembre.

2 http://eslo.huma-num.fr

corpus un modèle obtenu par le fine-tuning de FlauBERT décrit dans (Colli et al, 2025)[3]. Le schéma d'annotation utilisé suit l'approche de (Gosselin, 2010) et rend compte des 4 valeurs sémantiques de pouvoir selon notre approche : éventualité, possibilité matérielle et capacité, sporadicité et permission.

- **Étape 2. Tests statistiques** : en premier lieu, nous avons mené un test χ2 d'indépendance qui nous a permis de tester l'indépendance des variables "corpus" (13_11, ESLO_ENT, TWEETS, T2K_PEDIA) et "valeurs sémantiques *pouvoir*" (poss_mat_cap, éventualité, permission, sporadicité). En second lieu, afin de tester les hypothèses H1 et H2, nous avons mené un test post-hoc (comparaisons par paires avec ajustement de Bonferroni. 6 paires en total) et nous avons calculé la taille d'effet de chaque paire (coefficient V de Cramer). Comme expliqué dans (Wei et al., 2019), si le test du χ2 permet de savoir si les deux variables sont liées, il ne donne pas d'information sur l'intensité de cette relation. En outre, la significativité du résultat (p-value) peut être influencée par une taille élevée de l'échantillon.

Résultats

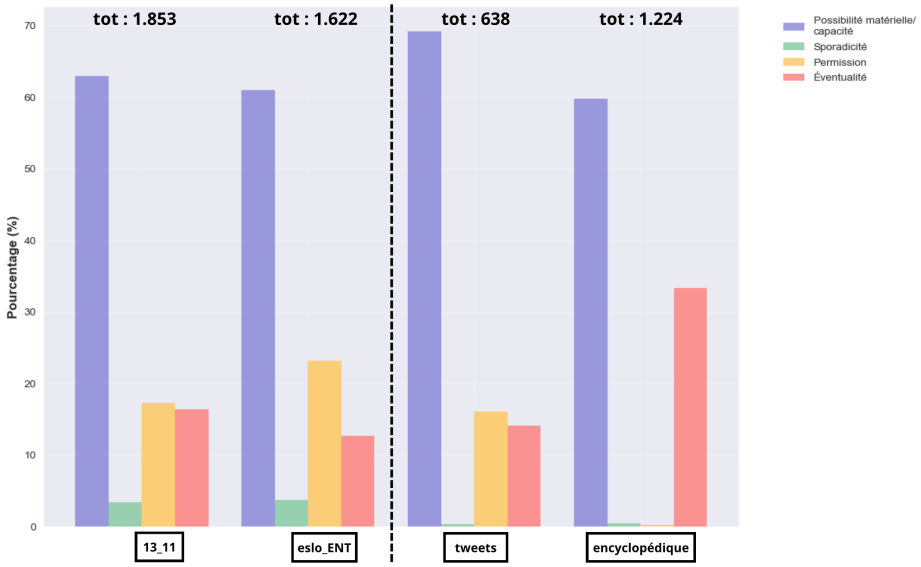


figure . 1 Proportion des valeurs sémantiques de *pouvoir* dans les quatre corpus.

Corpus1	Corpus2	Chi2	Bonferroni-adjusted p-value
ESLO_ENT	TK2_PEDIA	392.05	6.995578e-84
13_11	TK2_PEDIA	262.67	7.144104e-56
TWEETS	TK2_PEDIA	223.73	1.877749e-47
ESLO_ENT	TWEETS	35.51	5.693296e-07
13_11	ESLO_ENT	28.94	1.380357e-05
13_11	TWEETS	20.37	8.540514e-04

table 1. : Comparaison par paires

3 macro-avg = 0.92, entrainé et testé sur le volet "entretiens" du corpus eslo

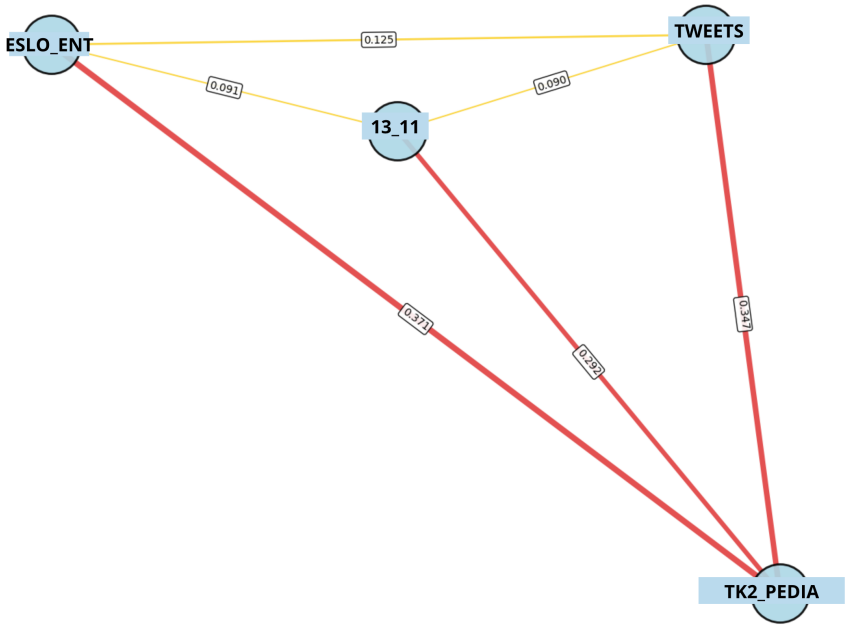


figure . 2 Proportion des valeurs sémantiques de *pouvoir* dans les quatre corpus.

Les résultats du test $\chi 2$ (X-squared = 473.48, df = 9, p-value < 2.2e-16) confirment que l'emploi des valeurs sémantiques de *pouvoir* varie de manière significative dans les 4 corpus. Dans la Figure 1 sont données à voir les proportions des différentes valeurs sémantiques de pouvoir dans les corpus. Le Tableau 1 présente les résultats de la comparaison entre paires (avec ajustement de Bonferroni). La Figure 2 montre les résultats du calcul du coefficient V de Cramer pour chaque paire. Le résultat indiqué sur chaque arête représente la taille d'effet de la corrélation entre la variable "modalité" et la variable "corpus", dont les valeurs correspondent, à chaque fois, aux deux nœuds situés aux extrémités de l'arête. Les valeurs du coefficient V de Cramer sont comprises entre 0 et 1. Plus elles se rapprochent de 1, plus la corrélation entre les deux variables est forte, ce qui traduit une plus grande divergence entre les distributions des valeurs sémantiques de pouvoir dans les deux corpus formant chaque paire. Selon le tableau proposé par (Cohen, 1988) valeurs du coefficient V de Cramer peuvent être interprétées comme suit : un effet négligeable pour V compris entre 0 et 0.10, un effet faible pour V compris entre 0.10 et 0.30, un effet moyen pour V compris entre 0.30 et 0.50, et un effet fort pour V supérieur ou égal à 0.50. En ce qui concerne (H1), nous remarquons une différence significative de distribution des valeurs sémantiques entre les deux corpus oraux (p-value = < 0.0001) avec un emploi statistiquement majeur des valeurs de sporadicité et d'éventualité dans le corpus 13_11. En ce qui concerne (H2), les résultats du Tableau 1 montrent, pour chaque paire, une différence significative dans la distribution des valeurs sémantiques de *pouvoir*. En revanche, les coefficients V de Cramer (Image 2) indiquent des tailles d'effet différentes selon le pair. Nous allons maintenant analyser en détail les différentes tailles d'effet. La comparaison de TWEETS et T2K_PEDIA avec chacun des deux corpus oraux donne des résultats différents. En effet, si la taille d'effet est moyenne pour les paires T2K_PEDIA - ESLO_ENT (0.4) et T2K_PEDIA -13_11 (0.3), ce n'est pas le cas de TWEETS. En effet, les tailles d'effet des paires TWEETS - 13_11 (0.09) et TWEETS - ESLO_ENT (0.1) sont entre négligeable et petite. Finalement, la taille d'effet pour les deux corpus écrits (T2K_PEDIA - TWEETS) est moyenne (0.4). Cela montre que l'emploi de pouvoir dans le corpus TWEETS est plus proche de celui observé dans des corpus oraux que de celui observé dans un corpus écrit comme T2K_PEDIA. En revanche, une analyse qualitative des résultats permet d'identifier des usages différents des mêmes valeurs sémantiques à l'intérieur des corpus. Par exemple, pour le pouvoir de permission, les tweets montrent surtout des permissions liées à des lois sociales (5), tandis que

les corpus oraux présentent plutôt des formes de politesse (6) ou des expressions typiques de l'oral (7) (comme "on peut dire")

5. Dans l'indifférence générale, à propos de #balancetonporc : "Bientôt on ne pourra plus dire bonjour à une fille." (TWEETS)

6. Euh attends j'ai un train de retard tu peux répéter ? (ESLO - ENT_1235)

7. Enfin j'ai fait essentiellement des mesures on peut dire [...] (ESLO - ENT_1014)

Bilan

Du point de vue de la distribution des valeurs sémantiques de pouvoir, les résultats obtenus ici(dé)montrent qu'elle change selon l'opposition oral vs. écrit. Les résultats (dé)montrent également qu'un corpus de communications médiées tel que le corpus TWEETS, bien que considéré comme du type écrit, s'écarte significativement de l'autre corpus écrit observé ici, TK2_PEDIA, mais se rapproche de corpus oraux comme 13_11 et ESLO_ENT. L'analyse qualitative des résultats offre un début d'explication à cette proximité.

Références bibliographiques

Barbet, C. (2012) Devoir et pouvoir, des marqueurs modaux ou évidentiels ? Langue française, 173, 49–63.

Battistelli, D. et Étienne, A. (2025). Modale, une ressource lexicale de la modalité au prisme des émotions. Carnets du Cediscor, numéro Marqueurs modaux, énonciation et argumentation, 35.

Chiril, P., Moriceau, V ., Benamara, F., Mari, A., Origgi, G. et Coulomb-Gully, M. (2020). An annotated corpus for sexism detection in French tweets. Proceedings of the Twelfth Language Resources and Evaluation Conference, 1397–1403, Marseille, France.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Colli, A. et Battistelli, D., Chagnoux, M. (2025). Quel usage des marqueurs modaux dans les discours post-traumatiques ? Carnets du Cediscor, numéro Marqueurs modaux, énonciation et argumentation, 35.

Colli, A. et Battistelli, D. (2025). Exploration de la modalité en français parlé et écrit. 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2025), Marseille, France.

Colli, A., Rossini, D. et Battistelli, D. (2025). A modal sense classifier for the French modal verb pouvoir. In Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), 233–243, Pisa, Italy.

Gosselin L. (2010). Les modalités en français : la validation des représentations, volume 1. Leiden, The Netherlands : Brill.

Pamuksaç, A. (2023). Quelle sémantique pour les verbes modaux du français? Étude des propriétés combinatoires de pouvoir, devoir, falloir et vouloir. Communication dans 11e Journées Internationales de Linguistique de Corpus (JLC), Grenoble, France.

Vetters C. (2012) Modalité et évidentialité dans pouvoir et devoir : typologie et discussions. Langue française, 173, 31–47.

Wei, R., Hu, Y., & Xiong, J. (2019). Effect Size Reporting Practices in Applied Linguistics Research: A Study of One Major Journal. *SAGE Open*, *9*(2).

Bilingual Language Use at the Dinner Table: A Corpus-Based Study of English–Russian Families in London

Yana Kut[1] et Aliyah Morgenstern[1]
[1] Laboratoire PRISMES, Université Sorbonne Nouvelle
yana.kut@sorbonne-nouvelle.fr, aliyah.morgenstern@sorbonne-nouvelle.fr

Introduction

In a world where nearly 43% of the population is bilingual, the dynamics of language use in multicultural cities like London offer a compelling lens into linguistic diversity and maintenance. According to the 2021 Census for England and Wales, just 0.4% of Londoners—around 32,000 people—report Russian as their primary language. Yet this figure likely underrepresents a much broader Russian-speaking community, including residents from former Soviet states. Amid this demographic complexity, sustaining Russian as a minority language in a predominantly English-speaking environment presents significant challenges—particularly in the absence of strong institutional support.

The present study will cover the topic of language dominance, and the family language practices in English-Russian bilingual families living in London. The research builds on the foundations laid by the ANR project DINLANG[1] (Morgenstern et al. 2021; Morgenstern & Boutet, 2024) and its comparison of *languaging*[2] practices in French speaking and French signing families during family dinners.

Previous studies have highlighted the complexity of language maintenance and loss when input in one language declines—particularly when only one parent speaks it and institutional support is lacking (De Houwer, 2007, 2015; Fishman, 1991; Hoff, 2012; Montrul, 2008; Tse, 2001). These findings underscore the importance of rich and varied input in the minority language to support its maintenance in bilingual children.

More recent research has increasingly examined the strategies used to maintain Russian as a heritage language in bilingual families living outside Russia, with a particular focus on the challenges its speakers face. This growing body of research spans countries such as Cyprus, Ireland, Israel, and Sweden (Karpava, 2021; Otwinowska et al., 2021), Germany (Brehmer, 2021), the United States (Wilkinson, 2024), as well as multi-country studies (Kisselev et al., 2024; Zabrodskaja & Ivanova, 2021). These works highlight the complexities of maintaining a weaker home language when it differs from the dominant language used in education, shaping how bilingual families navigate language use in everyday life.

This study examines how bilingual English-Russian families in London negotiate their language practices and how the dominant societal language, English, influences their children's

1 Agence Nationale de la Recherche, https://dinlang.ortolang.fr/

2 The authors adopt the term languaging (Linell 2009: 274) to include the different modes of expression such as speaking, signing, and gesturing.

bilingual development. It aims to reveal the relationship between different patterns of parental language input at home and children's use of the minority and majority languages across different ages.

Our theoretical framework is the usage-based theory of language acquisition (Tomasello, 2003), which posits that linguistic development is grounded in language use in context, shaped by frequency, function, and interactional experience (Bybee, 2010). From this perspective, language learning is driven by exposure to meaningful input and active participation in communicative practices. To capture the mechanisms of bilingual language practices more accurately, it is essential to analyse language use in natural settings (Han et al., 2024). Daily interactions, such as family meals, offer rich, recurring contexts in which children are exposed to and engage with language in functionally meaningful ways. These routine activities provide valuable insights into the development of linguistic competence.

Corpus and method

Corpus

To address these questions, we used the recordings of family dinners of three families from *The Kut corpus*. This corpus includes data from 11 bilingual families with target children aged between 2;7 to 13;4 at the time of the recording. The children are considered simultaneous bilinguals, having been exposed to both English and Russian from birth. The corpus, representing some 12 hours of natural interaction, was collected in London between October 2022 and February 2025 where one or two the family dinners as well as play sessions for 2 children were recorded at the family's home. All the participants signed an informed consent to participate in the study and filled in a questionnaire on the family's language practices and the protocol of the data collection was submitted and approved by the Sorbonne Nouvelle Ethics Committee.

Method

The recordings were transcribed in CLAN and coded in ELAN, with a specific focus on language choice, interlocutors and gesture productions to focus on all semiotic resources used in a multiparty framework.

This study focuses on a detailed analysis of three families from the corpus to illustrate the range of bilingual *languaging* observed depending on the families' linguistic choices. By closely examining these case studies, the aim is not to generalise, but to shed light on the nuanced and context-dependent dynamics that emerge in everyday language use across a variety of bilingual families' language practices.

For each family, two dinners were recorded 8 to 10 months apart, with no changes in context (same participants, same location), allowing us to analyse changes in the distribution of languages used by the bilingual children over time. The mothers in the first two families were raised monolingual in Russia or Belarus and started to acquire English sequentially at school and later after moving to London. The fathers in these families are native English speakers and only speak English to the children and their spouses. The one parent – one language strategy is used in both families when addressing the children. The third family is different in that both parents come from Belarus and were raised as monolinguals in Russian before moving to England; this is the language they use to talk to each other and to their children.

Our quantitative analysis of the children's language distribution during the family dinners recorded allows us to represent the children along a continuum¬—using the concept primarily developed by François Grosjean (Grosjean, 2015)—ranging from those with minimal proficiency in Russian to those who are nearly fluent.

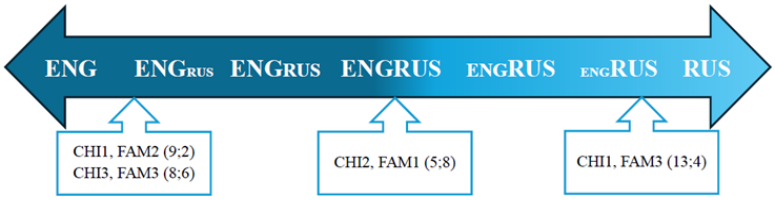


figure . 1 Bilingual continuum

Results

Thus, the youngest child in FAM1 can be said to illustrate the midpoint of the continuum occupied by the so-called balanced bilinguals whose language competencies are well developed in both languages, while the oldest children in FAM2 and FAM3 and the youngest in FAM3 are situated close to one of the extremes of the continuum, falling in the category of dominant bilinguals who are more proficient in one of their two languages. In the case of balanced use of both languages, the data confirm that the language choice of the target child in FAM1 (Robin, 5;8) depended primarily on the addressee: he spoke only Russian to his mother and only English to his father. In contrast, in FAM2, where both children (Oliver, 9;2, and Ian, 6;7) showed a clear dominance in English, their utterances were in English regardless of whether they were addressing their mother—who spoke to them in Russian—or their father.

A closer look at the data by conducting detailed qualitative analyses also allows us to observe a parallel tendency regarding the use of code-switching. Indeed, there seems to be a shift from an almost balanced use of the two languages in FAM1 and an absence of code-switching during the first dinner recorded to a more considerable use of the dominant language, English, and the appearance of mixed Russian-English utterances during the second dinner.

As for FAM3, code-switching is extensively used in the family by the parents and their three daughters, especially the two youngest. All family members interact using a blend of both languages, enriched by postures, gestures, and facial expressions that are fully integrated into their *bilanguaging*. This fluid, multimodal form of language practice is the mode they have naturally adopted for interacting with one another.

Thus, in accordance with the findings already provided by previous research, the data suggest that consistent and almost exclusive use of the minority language, Russian, by both parents as well as frequent use of code-switching, and additional input in Russian may be helpful in preserving the weaker language competences. Besides, birth order may also influence bilingual language development in favour of the heritage language; for example, being the firstborn—thus spending early, pre-school years without siblings—can result in increased exposure to the weaker language at home, potentially limiting the use of the dominant language by delaying peer influence and language shift.

Our initial hypotheses are supported by findings from Alice Brunet's research (2021) on French-English bilingual children in London and Paris, as well as preliminary quantitative data from a French–Russian family in *The Kut corpus*. Together, these results suggest that our observations may extend beyond the specific language pair studied.

Bibliography

Brehmer, M. (2021). Heritage language maintenance among Russian-speaking immigrants in Germany: Challenges and strategies. Journal of Language and Cultural Education, 9(3), 45–58.

Brunet, A. (2021). Acquisition and use of past tense-aspect morphology by French monolingual and French-English bilingual children. (Doctoral dissertation). Université Sorbonne Nouvelle.

Bybee, Joan (2010). Language, Usage and Cognition. Cambridge: Cambridge University Press.

De Houwer, A. (2015). Harmonious bilingual development: Young families' well-being in language contact situations. International Journal of Bilingual Education and Bilingualism, 18(6), 681–695.

Fishman, J. A. (1991). Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages. Multilingual Matters.

Grosjean, F. (2015). Parler plusieurs langues : Le monde des bilingues. Albin Michel.

Hoff, E. (2012). Language development. Wadsworth Cengage Learning.

Karpava, S., & Eracleous, A. (2021). Heritage language maintenance and revitalization: Evidence from Finland, Sweden, and Cyprus. National Heritage Language Resource Center.

Kisselev, O., Laleko, O., & Dubinina, I. (Eds.). (2024). Russian as a heritage language: From research to classroom applications. Routledge.

Montrul, S. (2008). Incomplete Acquisition in Bilingualism: Re-examining the Age Factor. Amsterdam: John Benjamins.

Morgenstern, A., Boutet, D. (2024). The Orchestration of Bodies and Artifacts in French Family Dinners. In Thiemo Breyer, Alexander Matthias Gerner, Niklas Grouls, Johannes F.M. Schick (Eds.) Diachronic Perspectives on Embodiment and Technology, pp.111-130, Springer.

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 L. Caronia (Ed.), Language and Social Interaction at Home and School (pp. 45–86). John Benjamins Publishing Company.

Otwinowska, A., Meir, N., Ringblom, N., Karpava, S., & La Morgia, F. (2021). Language and literacy transmission in heritage language: evidence from Russian-speaking families in Cyprus, Ireland, Israel, and Sweden. Journal of Multilingual and Multicultural Development, 42(4), 357–382.

Tomasello, M. (2003). Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge, MA: Harvard University Press.

Wilkinson, S. (2024). Revitalizing Russian: Heritage language education in the United States. Journal of Slavic Linguistics, 32(1), 112–130.

Zabrodskaja, A., & Ivanova, O. (2021). The Russian language maintenance and language contacts. Russian Journal of Linguistics, 25(4), 828–854.

C'est pas vrai, es stimmt nicht : les phrases préfabriquées à l'aune de la phraséologie contrastive

Alexia Jingand[1], Agnès Tutin[2]
[1] Laboratoire ATILF, Université de Lorraine
[2] Laboratoire LIDILEM, Université Grenoble Alpes
Alexia.jingand@univ-lorraine.fr, agnes.tutin@univ-grenoble-alpes.fr

Introduction

La linguistique de corpus ne contribue pas seulement au repérage et à la modélisation phraséologique, mais aussi à une approche plurilingue des phrases préfabriquées des interactions, désormais PPI (Tutin, 2022, 2019). Cette communication s'intéresse à deux couples de phrases relevant en français d'une même construction globale [c'est pas ADJ], avec la variable syntaxique [pas ADJ] : « c'est pas vrai » / « pas vrai » et « c'est pas possible » / « pas possible » et leurs traductions en allemand. Dans la perspective des grammaires de constructions (Deppermann, 2006 ; Goldberg, 2003), l'analyse envisagée recense des éléments lexicaux, syntaxiques, sémantiques et pragmatiques, notamment en consultant la structure interactionnelle et les fonctions langagières remplies par ces PPI. Ces critères constituent la clé de voûte de l'approche contrastive français-allemand, avec pour point de départ les choix traductologiques attestés dans des corpus alignés de traductions. Aussi l'intervention tâchera-t-elle de répondre aux questions suivantes : comment les constructions interactionnelles [c'est pas ADJ] sont-elles traduites en allemand ? Quels critères constructionnels (fonction, forme) sont priorisés lors du processus traductif de ces phrasèmes et lesquels s'effacent d'une langue à l'autre ? Comment la linguistique de corpus contribue-t-elle à une modélisation contrastive des phrases préfabriquées des interactions ?

Corpus et méthodologie

Corpus

Les corpus exploités relèvent du projet PREFAB (2022-2026). Parmi l'ensemble des corpus du projet, cette proposition se concentre sur les corpus d'écrits romanesques, dans lesquels se trouve de l'oral représenté, soit du discours direct (Rossi-Gensane, Etienne, Baldauf-Quilliatre, 2023, p. 48). Il s'agit donc d'interactions mises en scène dans la fiction, qui correspondent partiellement aux usages attestés dans l'oral spontané. Les données exploitées pour cette communication se déclinent comme suit :

Catégorie	Corpus monolingue	Corpus bilingues	
Nom du corpus	Phraséorom DE	Corpus aligné français-allemand	Corpus aligné allemand-français
Nombre d'œuvres	710	75	16
Nombre de tokens	86, 2 M	5, 7 M	0, 94 M

table 1. : Corpus du projet PREFAB exploité pour cette recherche

Le corpus principalement exploité pour cette étude constitue le corpus aligné du français vers l'allemand, tandis que le corpus monolingue en allemand et le corpus aligné de l'allemand vers le français permettront de mettre en regard les résultats obtenus et de comparer l'allemand langue source de l'allemand langue cible. Le corpus allemand langue-source intègre ainsi des données de vérification par rapport aux résultats obtenus dans le corpus aligné français-allemand. Le corpus a été formaté pour une exploitation à partir de la plateforme Lexicoscope (Kraif, 2016). Les œuvres ont été publiées de 1980 à nos jours, permettant d'aborder les données dans une approche synchronique.

Méthodologie

Nous observerons les stratégies traductionnelles mises en œuvre, à partir des différentes fonctions des PPI et leurs traductions en allemand. S'agit-il simplement de conserver la fonction pragmatique, sans recourir à une expression comparable dans la langue cible (Exemple : « c'est pas vrai » → « es stimmt nicht ») ou bien observe-t-on une grande proportion de PPI correspondantes en allemand (Exemple : « pas vrai » → « nicht wahr »). Les structures syntaxiques tendent-elles à être comparables (Exemple :[Pro Verbe Adj]) et le matériel lexical apparaît-il proche ? Après avoir observé ces stratégies, nous les confronterons aux corpus monolingues.

Ainsi, la méthodologie intègre une lecture empirique, guidée par les données :

1. Extraction et modélisation constructionnelle des couples de PPI « c'est pas vrai » / « pas vrai » et « c'est pas possible » / « pas possible » dans les corpus de français langue-source,

2. Extraction des traductions proposées pour ces deux couples de PPI en allemand langue-cible, analyse fréquentielle et constructionnelle des traductions, par exemple « das stimmt nicht »,

3. Extraction des traductions obtenues le plus fréquemment lors de l'étape (2) dans les corpus en allemand langue-source (Phraséorom et corpus aligné).

Cette méthodologie repose sur des distinctions qu'il conviendra d'expliciter, notamment autour des enjeux qui pose la notion d'équivalence en linguistique contrastive. L'analyse des résultats obtenus priorise l'approche par fonction pragmatique, une PPI renvoyant à diverses fonctions selon son contexte. Par exemple, « c'est pas vrai » se comprend à la fois comme une demande de confirmation, une remise en question de l'énoncé, ou encore une expression d'incrédulité.

Résultats

L'approche fait état d'une forte diversité lexicale dans les choix d'équivalence des PPI, ce qui indique une priorisation pragmatique, indicielle du « troisième code » auquel correspondrait la traduction (Baker, 1998). Pour la recherche en phraséologie, les corpus alignés permettent ainsi de mettre en évidence les fonctions pragma-sémantiques de certaines PPI et de compléter ou affiner la modélisation monolingue. « Pas possible » est par exemple traduit par « unmöglich » (*pas possible*) mais aussi par « unglaublich » (*pas croyable*) ou encore par l'interjection polylexicale interrogative « Ach was? » (*ah bon ?*). Nous observerons dans quelle mesure les phrases préfabriquées se dissolvent dans le processus traductif, où le degré de figement cède la priorité à la traduction de la fonction pragmatique.

Du point de vue des fonctions langagières, on remarque que « C'est pas vrai » est traduit essentiellement par « das ist nicht wahr », dont les apparitions en allemand langue-source ne

correspondent pas aux mêmes fonctions. Il ne s'agit plus de mise en cause d'un énoncé ou d'une expression d'incrédulité, mais d'une demande de confirmation (Gipper, König & Weber, 2024). Une équivalence sémantico-pragmatique de « pas vrai » consiste par exemple en une construction d'une autre nature en allemand : « es stimmt nicht », plus proche des fonctions communicatives de l'énoncé-source, mais moins transparente d'un point de vue lexico-syntaxique.

Par ailleurs, la comparaison syntaxique des traductions des PPI verbales et averbales témoigne d'une tendance à rendre une construction verbale par une autre, tandis que les constructions averbales sont fréquemment traduites par des constructions verbales. Ainsi, la phrase « pas possible » se voit traduite par « unmöglich », mais aussi par « ist doch nicht möglich! », qui intensifie la proposition originale. À l'inverse, les PPI « c'est pas vrai » et « c'est pas possible » ne rencontrent jamais de traductions averbales.

Pour conclure, cette recherche poursuit les réflexions entreprises en phraséologie contrastive, et contribue à l'observation des critères d'équivalence (Schafroth, 2020) privilégiés par la traduction. Étayée de données chiffrées précises, il s'agira de cartographier les stratégies traductives utilisées pour les PPI, en observant un cas de convergence de structure, un cas de PPI traduite par une non PPI, et de formuler des hypothèses concernant les motifs qui expliquent le choix d'une stratégie plutôt qu'une autre. Les trajectoires méthodologiques de cette étude participent à déterminer la variation traductologique par opposition à l'hypothèse d'un « troisième code », comprise comme une « forme de communication [qui] subit des contraintes spécifiques dont les messages traduits portent la marque » (Kraif, 2017, p. 58), en se concentrant sur un couple de PPI pour lequel la linguistique contrastive met en exergue la richesse sémantico-pragmatique.

Références bibliographiques

Baker, M. (1998). Réexplorer la langue de la traduction : une approche par corpus. *Meta. Journal des traducteurs.* 43*(4)*, 1-7. https://doi.org/10.7202/001951ar

Deppermann, A. *Construction Grammar* – Eine Grammatik für die Interaktion? Deppermann, A., Fiehler, R. & Spranz-Fogasy, T. (dir.) *Grammatik und Interkation.* Radolzfell: Verlag für Gesprächsforschung, 43-65.

Gipper, S., König, K. & Weber, K. (2024). Structurally Similar Formats Are Not Functionally Equivalent across Languages: Requests for Reconfirmation in Comparative Perspective. *Contrastive Pragmatics*, 5, 195-237.

Goldberg, A. (2003). Constructions: a new theoretical approach to language. *TRENDS in Cognitive Sciences*, 7*(5)*, 219-224. https://doi.org/10.1016/S1364-6613(03)00080-9

Kraif, O. (2017). Traduire le polar : une étude textométrique comparée de la phraséologie du roman policier en français langue source et cible. *Synergies Pologne*, 14, 43-60.

Kraif, O. (2016). Le Lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arbordés. *Cahiers de Lexicologie*, 108*(1)*, 91-106.

Schafroth, E. (2020). Why Equivalence of Idioms in Different Languages is the Exception. Arguments from a Constructional Perspective. Cotta Ramusino, P. & Mollica, F. (éd.). *Contrastive Phraseology: Languages and Cultures in Comparison*. Cambridge Scholars Publishing, 129-150.

Rossi-Gensane, N., Etienne, C. & Baldauf-Quilliatre, H. (2023). La phrase préfabrique *c'est ça* dans les interactions réelles et représentées. *Studii de lingvistică,* 13*(2)*, 35-60.

Tutin, A. (2022). *Comment dirais-je ? Comment veux-tu ? Comment ça va ?* Quelques observations sur les phrases interrogatives partielles préfabriquées dans les interactions orales et les dialogues romanesques. *Lingvisticae Investigationes*, 45*(2)*, 172-196.

Tutin, A. (2019). Phrases préfabriquées des interactions : quelques observations sur le corpus CLAPI. *Cahiers de lexicologie*, 114, 63-91.

Caractéristiques de l'interaction dans les SMS vocaux en France et au Québec : les formules d'ouverture et de clôture

Julie Glikman [1], Anne-Sophie Bally [2]
[1] Université de Lorraine, CNRS, ATILF, F-54000 Nancy, France
[2] Université du Québec à Trois-Rivières, Trois-Rivières (Québec), Canada
julie.glikman@univ-lorraine.fr, Anne-Sophie.Bally@uqtr.ca

Introduction

Les formules d'ouverture et de clôture de l'interaction, en particulier les formules de salutation, sont constitutives des interactions, dont elles participent à la construction et l'élaboration. Elles sont de ce fait centrales dans l'analyse des conversations et sont particulièrement représentatives des variations culturelles des règles conversationnelles (Kerbrat-Orecchioni, 2016). Elles restent cependant encore difficiles à documenter en contexte naturel, vu les difficultés à réunir massivement des données authentiques produites hors de tout contexte de recherche et en milieu informel. Le TLFi, par exemple, donne *salut* comme familier et ne mentionne même pas l'emploi de *coucou* comme formule de salutation, qui s'est pourtant largement répandu en français hexagonal. Wikipédia, sous *salutation* (consulté le 16.04.25), donne *salut* comme « Salut francophone familier » signifiant aussi bien *bonjour* que *au revoir*, et synonyme de *coucou*, sans précision géographique. La variante *bon matin*, répandue au Québec, n'est quant à elle décrite dans aucune des deux sources, et mentionnée principalement dans des chroniques (le plus souvent pour critiquer ce calque de l'anglais, cf. cet article de La Presse du 22 août 2021 consulté le 16.04.25). L'analyse de deux corpus de données authentiques de SMS vocaux récemment constitués, *Les Vocaux* et *Sors ton vocal* (dont les caractéristiques sont détaillées ci-dessous dans la section Corpus), devrait permettre de combler ce manque. L'objectif de cette contribution est de fournir une analyse qualitative des formules d'ouverture et de clôture dans ces deux corpus de SMS vocaux. En tant que constitutives de l'interaction, nous postulons que l'étude de ces formules contribue à la caractérisation de nos données, et participe à la description fine de ce nouveau mode de communication médié. Les vocaux sont en effet une pratique émergente qui reste encore à décrire et à définir du point de vue communicationnel. Dans la mesure où la constitution d'un corpus de SMS vocaux implique un mode de collecte nécessitant assez peu de ressources humaines (l'enregistrement est pris en charge par le participant lui-même), les linguistes pourraient vouloir l'utiliser pour constituer d'autres corpus en ciblant des populations (ex. la parole des jeunes). Nous pensons toutefois qu'il faut dès maintenant caractériser ces données pour bien cerner le type d'information qu'on peut en extraire. De plus, les SMS vocaux appartiennent à la sphère privée, ce qui les rend particulièrement susceptibles d'être représentatifs des variations culturelles. L'analyse de ces formules nous permettra de mettre au jour les points de convergence et de divergence entre les variétés du français d'Europe (Les Vocaux) et du Québec (Sors ton vocal).

Corpus et méthodologie

Corpus

Notre étude s'appuie sur deux corpus de SMS vocaux, récoltés de manière indépendante en France (corpus Les Vocaux, constitué dans le cadre du projet Oralidia) et au Québec (projet Sors ton vocal!). Dans les deux cas, les SMS vocaux récoltés dans ces corpus sont des SMS authentiques, qui ont été réellement produits et envoyés en dehors des besoins de la collecte. Les données formelles de chaque corpus sont détaillées dans la table 1. Le mode de recueil n'a cependant pas permis de collecter les messages dans leur contexte interactionnel. Les SMS vocaux constituent un nouveau mode de communication médiée, en lien avec l'évolution technologique. Ils présentent des caractéristiques hybrides du point de vue communicationnel, en particulier en suivant la grille d'analyse de Koch et Oesterreicher 2001 (voir notamment Mazziotta et Glikman 2023 ; Glikman et Fauth 2022). Formellement, ils constituent des monologues : il n'y a pas d'interruption possible, pas de chevauchement de parole, pas de co-construction du discours. Ils ne relèvent donc pas de la conversation telle que définie par Kerbrat-Orecchioni (2016, p. 123) : « toute conversation est une construction collective ». Cependant, ils sont bien adressés à un destinataire (unique ou multiple) identifié et connu, mais absent. En tant que communication médiée, les vocaux présentent un éloignement spatial et temporel (même si le rythme des échanges donne parfois une illusion de simultanéité). Ce sont des communications différées et asynchrones, médiées par un support technologique.

Corpus (et son abréviation)	Les Vocaux (LV)	Sors ton vocal (STV)
Années de collecte	2021-2022	2024-2025
Nombre de participants	45	91
Origine géographique des participants	France principalement, Belgique, Suisse	Résidents de la province de Québec (Canada)
Âge des participants	De 11 ans à 44 ans	De 14 ans à plus de 70 ans
Nombre total de SMS vocaux recueillis	1160	953
Durée totale	19h06m49.2s	14h07m21s
Équivalent en nombre de mots transcrits	Env. 255 000	En cours de transcription

table 1. : Caractéristiques des corpus *Les Vocaux* et *Sors ton vocal*

Méthodologie

Pour mener à bien cette étude, nous avons relevé dans nos corpus les formules d'ouverture et de clôture des messages, c'est-à-dire les éléments se trouvant directement au commencement de l'enregistrement et à la fin de l'enregistrement. Les corpus étant encore en cours de constitution, les résultats présentés dans la section ci-dessous ne contiennent pas de mesures quantitatives, mais donnent à voir les principales tendances relevées dans nos corpus.

Résultats

L'observation de nos données nous permet dans un premier temps de distinguer deux types de messages, selon la présence ou l'absence de ces formules. Certains vocaux ne comportent en effet pas de formules d'ouverture ou de clôture de type salutation : « alors au niveau des cours euh je faisais vraiment [...] comment ça se fait qu'il y a eu différentes langues qui sont descendues du latin en fait » (LV 433_85). Deux hypothèses peuvent expliquer cette absence : il s'agit soit d'un message interne à une interaction (notamment dans les messages contenant une interrogation directe), soit d'une possibilité de passer outre ces rituels conversationnels dans ce type de communication, ce que laissent penser les témoignages lors des passations

d'entretien, les participants arguant que c'est plus facile de couper la communication, qu'on n'a pas besoin de dire qu'on arrête la conversation, avec les SMS vocaux (par rapport à l'appel téléphonique notamment). En lieu et place de la salutation, on trouve d'autres types de marqueurs, qui peuvent servir de marques d'ouverture, comme ici *alors*, mais également *ouais, en fait, donc, du coup, après* (LV) ou *pis/et puis, fait que, wow, oh là là, mais là, ben en fait, ben non* (STV) et qui peuvent aussi être la trace d'une interaction en cours – même si elle est asynchrone. En effet, certains de ces marqueurs indiquent une réaction à des propos antérieurs ou une rectification de ceux-ci, comme *ben non* dans *Ben non, elle est pas vendue* [STV_032_20241227_02]. D'ailleurs, l'emploi du pronom anaphorique *elle* en ouverture, sans réintroduction du référent, est aussi une marque d'interaction, mais son étude dépasse les limites de cette communication. La faible présence de formules conclusives sur le temps bref comme *à plus (tard), à très vite, à tantôt, à tout à l'heure, à tout de suite, à toute* laisse à penser que l'émetteur du message n'anticipe pas un retour ou une réponse. Dans le questionnaire STV, les participants ont massivement indiqué utiliser le vocal pour éviter d'entrer en dialogue avec le récepteur (p. ex. pour se soustraire à une longue discussion ou pour ne pas déranger l'autre), ce que l'absence de ces formules semble confirmer.

Dans le cas des vocaux qui comportent des formules d'ouverture et/ou de clôture, leur forme permet de situer l'interaction sur une échelle de degré de familiarité avec le destinataire, par exemple l'emploi de *coucou* vs *bonjour* dans la variété européenne (LV, cf. loc. 106) et de relever des emplois émergents encore peu documentés, comme le terme *meuf* en apostrophe (LV, 70_35), tandis que la forme *salut*, pourtant donnée comme courante en français hexagonal, est peu représentée dans LV, mais bien implantée dans STV. Ces formules mettent également en évidence la variation régionale. En salutation d'ouverture, LV et STV partagent *coucou* et *yo*. LV contient en plus *recoucou, hello* et *hey* tandis que STV contient *allo* et *heille.* En salutation de clôture se trouvent *bisous, je t'embrasse, je t'aime* dans LV alors que *bye, bye bye* et *ok bye* apparaissent dans STV. S'observent également de possibles idiosyncrasies : dans LV, la locutrice 106 commence systématiquement par *oui*, en combinaisons diverses : *oui coucou* (LV 106_56), *oui bonjour* (LV 106_46), *oui bonsoir* (LV 106_50), ou encore *oui donc rebonjour* (LV 106_49). On remarque aussi en clôture des formules métadiscursives, du type "je finis dans un autre message" (LV 437_17) qui annoncent un autre message à venir, par opposition à des formules comme "j'attends de vos nouvelles", qui donnent la parole au récepteur du message (STV_014_20241203_04). En clôture, les données québécoises et françaises présentent souvent respectivement *c'est ça* et *voilà*, qui permettent à l'émetteur de dire que le message est complet.

En conclusion, malgré les limites de nos corpus (messages recueillis en dehors de leur situation interactionnelle), nos données nous permettent de caractériser les vocaux comme type de production, en particulier de confirmer que ces messages participent de l'interaction. Des études complémentaires sur les termes d'adresse, la présence des personnes de l'énonciation et la présence d'énoncés visant le destinataire (ordre, question directe) restent à faire. Les vocaux constituent des formes d'interaction spontanée et vernaculaire, non surveillée d'une manière générale (dans la limite de ce que suppose l'action même, cf. Mondada, 2015). Au sein des vocaux eux-mêmes, la prise en compte des formes d'ouverture et de clôture permet de faire des hypothèses sur la classification selon le degré de familiarité avec l'interlocuteur. Cette étude confirme la nécessité de continuer à travailler à la constitution de corpus écologiques pour l'analyse des productions langagières en contexte naturel, en contexte médié, voire en contexte humain-machine.

Références bibliographiques

ATILF - CNRS & Université de Lorraine. (1994). *Salut*. Trésor de la langue française informatisé (TLFi). http://stella.atilf.fr/Dendien/scripts/tlfiv5/visusel.exe?12;s=568212675;r=1;nat=;sol=1;

Bally, A.-S., Lévesque, I. et Vallerand, N. (à paraitre). Recueillir le français parlé en conversation privée sur les messageries numériques : enjeux éthiques et méthodologiques. Dans K. Reinke et W. Remysen (dir.), *Le français dans les médias : Pratiques langagières non standardisées, attitudes et représentations dans les formats médiatiques oraux associés au divertissement*. Presses de l'Université Laval.

Glikman, J. et Fauth, C. (2022). Un nouvel accès à la parole spontanée : les vocaux. *Proc. XXXIVe Journées d'Études sur la Parole - JEP 2022* (pp. 154-162). ISCA, doi: 10.21437/JEP.2022-17

Glikman, J., Mazziotta, N., Benzitoun, C., et al. (2025). *LesVocaux* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v0.0.2, https://hdl.handle.net/11403/lesvocaux/v0.0.2.

Kerbrat-Orecchioni, C. (2016). L'analyse des conversations. Dans J. Dortier La Communication : Des relations interpersonnelles aux réseaux sociaux (p. 122-129). Éditions Sciences Humaines. https://doi.org/10.3917/sh.dorti.2016.02.0122.

Koch, P. et Oesterreicher, W. (2001). Gesprochene Sprache und geschriebene Sprache / Langage parlé et langage écrit. Dans G. Holtus, M. Metzeltin et C. Schmitt (Dir.), *Lexikon der Romanistischen Linguistik, 1*(2) (pp. 584-627). De Gruyter. https://doi.org/10.1515/9783110938371.584

Mazziotta, N. et Glikman, J. (2023). Emplois discursifs et pragmatiques des formes du verbe écouter Observations sur les corpus 88milSMS et Les Vocaux. Dans M. Saiz-Sanchez et S. Gomez-Jordana Ferary (dir.), *Études de sémantique et pragmatique en synchronie et diachronie* (p. 23-42), Presses Universitaires Savoie Mont Blanc.

Mondada, L. (2015). Ouverture et préouverture des réunions visiophoniques. Réseaux, 194(6), 39-84. https://doi.org/10.3917/res.194.0039.

Morel, E. (2018). Textos : assemblages hétérosémiotiques : Approche plurielle des pratiques plurilingues dans la communication par SMS et WhatsApp. De Boeck Supérieur. https://doi.org/10.3917/dbu.morel.2018.01

Salutation. (2025, 16 avril). Dans Wikipédia. https://fr.wikipedia.org/wiki/Salutation

Comparaison des traces d'élaboration de textes narratifs oraux et écrits d'enfants francophones de 10-11 ans

Emilie Ailhaud [1] et Florence Chenu [2]
[1] UR CONFLUENCE : Sciences et Humanité (EA 1598), Université Catholique de Lyon
[2] Laboratoire Dynamique Du Langage (UMR 5596), Université Lyon2 & CNRS

Introduction

La production de textes à l'oral ou à l'écrit est une activité complexe et coûteuse cognitivement, particulièrement lorsque chez les individus, les processus qui les permettent sont au début de leur développement. La production de textes à l'écrit est réputée être moins impactée par la contrainte temporelle que la production à l'oral. Les scripteurs peuvent à l'écrit prendre du temps pour organiser leurs idées, choisir les structures linguistiques. Lors de productions orales, les contraintes communicatives peuvent à l'inverse amener une certaine pression temporelle aux locuteurs. Les recherches montrent que la production de textes à l'oral (par exemple en utilisant la dictée à l'adulte) chez les enfants scripteurs débutants aboutit à des textes plus longs, de meilleure qualité et comportant davantage d'idées qu'à l'écrit (Hayes et Berninger, 2010), et les textes écrits se distinguent des textes oraux au niveau lexical et syntaxique, notamment avec des structures plus fréquentes à l'oral (Jisa, 2004 ; Johansson, 1999). Dans cette étude, nous proposons de comparer des traces de l'élaboration de textes narratifs oraux et écrits d'enfants francophones de 10-11 ans (scolarisés au niveau CM2 en France) afin de mieux comprendre les différences entre la production à l'oral et à l'écrit.

Plusieurs éléments peuvent être analysés pour étudier les traces d'élaboration. Tout d'abord, les pauses et débits peuvent être des indices des processus cognitifs en cours. En effet, les ressources cognitives, et notamment la capacité de la mémoire de travail, sont limitées (McCutchen, 1996) : face à une demande cognitive trop coûteuse, l'individu qui est en train de produire un texte peut donc être amené à ralentir son débit de parole ou d'écriture, voire à s'arrêter, plus ou moins longuement. Des longues pauses sont ainsi particulièrement relevées en début d'unité syntaxique : à l'écrit, il a été fréquemment montré que plus l'unité était élevée dans la hiérarchie syntaxique (paragraphe, phrase), plus la pause était longue (voir entre autres Foulin, 1998; Schilperoord & Sanders, 1999). Cela peut amener à considérer que ces unités sont, à l'écrit, des unités de planification. Les pauses peuvent aussi être associées à d'autres processus rédactionnels, comme la révision. À l'oral, on peut relever des pauses silencieuses et des pauses remplies, notamment avec des mots du type « euh », et des allongements de voyelles (Fox, 2010). Ces pauses peuvent marquer une hésitation, en début ou en milieu d'unité, ou avoir plus clairement une fonction démarcative (Campione & Veronis, 2005). Les répétitions et faux-départs peuvent également témoigner de processus d'élaboration.

Comparer des traces d'élaboration dans des corpus de textes écrits et oraux fait émerger plusieurs enjeux. Tout d'abord, il convient de recourir à des outils de collecte de données et à un codage suffisamment précis pour relever les différents indicateurs – longueur des pauses, pauses remplies, reprises - dans chacune des modalités. De plus, une certaine rigueur d'analyse est nécessaire pour faire émerger les fonctions de ces différents indicateurs, afin d'identifier les éléments qui témoigneraient particulièrement du processus d'élaboration.

Corpus et méthodologie

Corpus

Pour cette étude, les données ont été collectées auprès de 40 participants francophones âgés de 10-11 ans dans des établissements scolaires de la région lyonnaise. Nous avons sollicité de la part de chaque individu la production de textes narratifs à l'oral et à l'écrit à propos du conflit entre les gens, thème qui s'était trouvé en effet être porteur lors de précédentes études et d'études pilotes (Gayraud, 2001). Après avoir visionné une vidéo présentant une série de scénettes sans parole portant sur des problèmes courants au milieu scolaire (triche, disputes, vol…), les enfants devaient raconter une histoire portant sur une dispute ou un problème qu'ils avaient pu avoir avec une autre personne.
La moitié des participants a produit en premier le texte narratif à l'oral puis le même texte à l'écrit et l'autre moitié des participants a commencé par produire le texte narratif à l'écrit, puis a produit le même texte à l'oral.

Les textes oraux ont été enregistrés, puis transcrits au format Childes (MacWhinney, 2000). Les textes écrits ont été collectés sur des tablettes graphiques associées au logiciel Eye & Pen © (Chesnet & Alamargot, 2005), permettant d'enregistrer le déroulement de l'écriture et exportés au format Childes.

Méthodologie

Cette étude répertorie dans les textes oraux et écrits les traces de leur élaboration. Ainsi, à l'oral ont été codées les pauses silencieuses, pauses remplies, répétitions, faux-départs, auto-corrections. À l'écrit ont été codées les pauses et les révisions.

La fréquence de ces différents phénomènes ainsi que leur localisation dans les unités linguistiques (début de phrase, début de proposition, autre) est examinée en fonction de la modalité (oral, écrit) et de l'ordre de passation (texte oral produit en premier ou texte écrit produit en premier).

Résultats

Nous faisons les hypothèses suivantes :

- Les textes oraux présentent proportionnellement davantage de traces d'élaboration ayant une contrepartie sonore (répétitions, reprises, pauses remplies), alors que les textes écrits présentent proportionnellement davantage de pauses.
- Les pauses sont plus longues à l'écrit qu'à l'oral.
- La localisation des traces d'élaboration respecte davantage les frontières d'unités linguistiques à l'oral qu'à l'écrit.
- Les traces d'élaboration sont plus nombreuses dans les textes produits en premier, que ce soit à l'oral ou à l'écrit.

Les résultats montrent que les pauses supérieures à 1 seconde sont proportionnellement plus nombreuses dans les textes écrits que dans les textes oraux. Les reformulations à l'oral sont proportionnellement plus nombreuses qu'à l'écrit – si on ne considère à l'écrit que les interruptions de mot suivies d'un nouveau mot, et non les révisions portant sur l'orthographe. Par ailleurs, les textes produits d'abord à l'écrit contiennent davantage de pauses, mais il n'y a pas de différence concernant les reformulations.

Cette étude permet dans un premier temps de clarifier les différents types d'indicateurs du processus d'élaboration, en tenant compte des spécificités de l'oral et de l'écrit et des contraintes de chacune de ces modalités. Dans un deuxième temps, les comparaisons en fonction de l'ordre de passation ouvrent à des questionnements d'ordre didactique, concernant le travail sur les textes, à la fois écrits et oraux. Ces questionnements pourront être poursuivis, dans des études ultérieures, par une exploration plus large de corpus à notre disposition, par exemple en comparant différents types de texte ou des textes produits pas des locuteurs de différents âges.

Références bibliographiques

Campione, E., & Véronis, J. (2005). Pauses and hesitations in French spontaneous speech. *DiSS, 5*, 43-46.

Chesnet, D., & Alamargot, D. (2005). Analyse en temps réel des activités oculaires et graphomotrices du scripteur : intérêt du dispositif « Eye and Pen ». *L'année Psychologique, 105(3),* 477–520.

Foulin, J. N. (1998). To what extent does pause location predict pause duration in adults' and children's writing? *Cahiers de Psychologie Cognitive, 17(3),* 601–620.

Fox, Barbara. 2010. Introduction. In Nino Amiridze, Boyd H. Davis & Margaret Maclagan (eds.), *Fillers, pauses and placeholders (Typological Studies in Language 93)*, 1–9. Amsterdam, Philadelphia: John Benjamins.

Gayraud, F. (2001). *Le développement de la différenciation oral/écrit vu à travers le lexique* (Thèse de doctorat). Université lumière-Lyon2, Lyon.

Hayes, J. R., & Berninger, V. W. (2010). Relationships between idea generation and transcription: How act of writing shapes what children write. In C. Braverman, R. Krut, K. Lunsford, S. McLeod, S. Null, P. Rogers, & A. Stansell (Eds.), Traditions of writing research (pp. 166–180). Taylor & Francis.

Jisa, H. 2004. Growing into academic French. Later Language Development: Typological and psycholinguistic Perspectives, Collection *Trends in Language Acquisition R*esearch (TILAR) 3: 135-162.

Johansson, V. 1999. Word frequencies in speech and Writing: a study of expository discourse. *Genre and modality in developing discourse abilities* 1: 182-198.

MacWhinney, B. (2000). *The CHILDES project: The database (Vol. 2).* Psychology Press.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational psychology review, 8*, 299-325.

Schilperoord, J., & Sanders, T. (1999). How hierarchical text structure affects retrieval processes: Implications of pause and text analysis. In *Knowing What to Write. Conceptual Processes in Text Production* (pp. 13–33).

Défis et enjeux dans la création d'un dispositif didactique pour des apprenti-es allophones à partir d'un corpus spécialisé

France Rousset[1]
[1] École de langue et civilisation françaises, Université de Genève
france.rousset@unige.ch

Introduction

Dans un contexte actuel caractérisé par une forte migration des populations, les apprenti-es migrant-es inscrit-es dans un cursus de formation professionnelle font face à plusieurs défis, notamment langagiers. En effet, ce public particulier doit apprendre la langue du pays d'accueil, ainsi que le langage propre au métier en cours d'appropriation (André et Adami, 2013). Au fil du temps, la part langagière du travail a gagné en importance dans tous les secteurs professionnels (Boutet, 2001). Ainsi, les apprenti-es allophones se doivent de développer des compétences langagières afin de pouvoir comprendre et participer à ces interactions professionnelles.

Dans la formation professionnelle, des études sur les activités interactionnelles entre formateur/trices et apprenti-es ont permis d'identifier et de décrire des moments significatifs sur le plan didactique, à l'exemple des consignes (Filliettaz, 2009). Ces séquences de consignes (Filliettaz, 2009 ; Veyrac, 2001), nombreuses en contexte d'apprentissage, représentent un des moments interactionnels hautement importants pour la réussite de la formation. Or, elles constituent un grand défi de compréhension pour l'ensemble des apprenti-es, et ce même en fin de formation (Rousset et Thomas, in *press*).

Dans cette contribution, nous présenterons une recherche doctorale en cours s'articulant autour du triptyque suivant : analyse interactionnelle et multimodale de séquences de consignes, didactique du français langue d'insertion et d'intégration (désormais FL2I) et formation professionnelle initiale. Dans cette optique, nous montrerons à partir d'analyses interactionnelle et multimodale de séquences de consignes du corpus audio-vidéo spécialisé – DiCoi pro[1], la complexité de ces séquences faisant intervenir plusieurs phénomènes liés à la compétence d'interaction. Nous réfléchirons ensuite à la manière de transposer ces observations en un dispositif didactique pour les apprenti-es allophones en formation professionnelle. Ceci nous permettra de présenter les défis et enjeux liés à la mise en place d'un tel dispositif visant le développement de la compétence d'interaction (Pekarek-Doehler, 2006) des apprenti-es allophones. Nous présenterons enfin les résultats attendus de ce dispositif et ses répercussions pour l'ensemble de la formation initiale.

1 Les données ont été récoltées par France Rousset dans le cadre du projet DiCoi mené à l'Institut de plurilinguisme (programme 2021-2024 du Centre scientifique de compétence sur le plurilinguisme) et sous la direction de Prof. Anita Thomas.

Corpus et méthodologie

Corpus

Dans cette étude, nous exploitons le corpus DiCoi pro que nous avons recueilli *in situ* dans les locaux de centres professionnels en Suisse romande lors d'un projet antérieur en didactique du FLE (Rousset et Thomas, in *press*). Les données analysées font partie de la collection « consignes » élaborée pour notre projet doctoral. Comme critère principal de constitution de cette collection, nous avons fait le choix d'arrêter les extraits à la frontière entre le *dire de faire* du formateur et le *faire* de l'apprenti-e. Cela nous a permis d'obtenir un total de 67 extraits provenant de quatre secteurs de formation (pâtisserie, confection, menuiserie, restauration). Dans ces données, nous avons identifié une hybridité au niveau des environnements avec des espaces à dominante éducative (p.ex. : atelier pâtisserie du centre de formation) et des espaces à dominante productive (p.ex. : service du midi dans la restauration au sein du restaurant rattaché au centre de formation). Tous les extraits comprennent des interactions de consignes entre Maitres socioprofessionnels (MSP) et apprenti-es, transcrites de manière multimodale dans le logiciel ELAN (Mondada, 2017).

Méthodologie

La mise en place d'un dispositif didactique correspond au deuxième volet de la thèse et découle directement du premier volet (analyse interactionnelle et multimodale des consignes). Ce dispositif est en cours d'élaboration pour une expérimentation prévue au printemps 2026.

Cherchant à décrire l'activité de consignes dans les interactions, nous avons d'abord analysé selon une approche interactionnelle et multimodale (Mondada, 2017) les extraits de la collection « consignes ». Les principaux résultats ont mis en lumière plusieurs phénomènes liés à la compétence d'interaction, à l'illustration de la séquentialité, de l'organisation spatio-temporelle de ces interactions ou encore de la co-construction de la consigne tout en soulignant le rôle central des gestes, devenant par moments partie constituante de la consigne. Le caractère fragmenté des consignes diffusées tout au long d'une même activité a permis d'établir une typologie des consignes (p.ex : consignes inaugurales, réparatrices, progressives).

En partant de la complexité de ces séquences interactionnelles fortement importantes et abondantes dans la formation, le dispositif didactique, en cours d'élaboration, vise le développement de la compétence d'interaction d'allophones en pré-apprentissage en Suisse romande. Les enjeux liés à ce dispositif sont multiples. En effet, il a pour ambition d'amener les apprenti-es allophones vers une meilleure compréhension des consignes, mais aussi vers l'appropriation de stratégies pour gérer des séquences interactionnelles complexes en formation professionnelle.

Nous avons envisagé un dispositif en deux parties. En implémentant l'approche de l'analyse interactionnelle dans un contexte encore peu étudié, celui de la didactique des langues en formation professionnelle, nous proposons quatre analyses collectives (André, 2019 ; Filliettaz, 2022) de séquences de notre collection « consignes » avec les apprenti-es allophones sur la période d'un semestre. Ces analyses collectives seront filmées et permettront ensuite d'étudier les paroles *sur* les interactions et les interactions *entre* les apprenti-es. En complément de ces analyses en classe, dix exercices interactifs permettront un travail plus approfondi sur les phénomènes de la compétence d'interaction.

La création du dispositif fait face à de nombreux défis didactiques tels que la familiarisation des apprenti-es allophones à la démarche des analyses collectives ; l'utilisation du métalangage et la granularité de guidage lors des co-analyses.

Lors de la phase d'expérimentation, il ne s'agira pas de mesurer la production des apprenti-es allophones, mais d'observer comment ils/elles s'engagent dans les analyses collectives tout au long du semestre.

Résultats attendus

Parmi les résultats attendus de l'expérimentation de ce dispositif, nous pouvons citer les trois suivants.

- Premièrement, grâce à l'écoute répétée d'extraits authentiques (dans les deux volets du dispositif), les apprenti-es acquièrent une meilleure compréhension des interactions de manière générale.
- Deuxièmement, par l'intermédiaire de co-analyses de situations authentiques, les allophones sont amenés à une prise de conscience du mécanisme des interactions professionnelles, avec un focus tant sur les ressources verbales que multimodales utilisées.
- Troisièmement, en partant d'extraits authentiques montrant la co-construction de consignes, les apprenti-es développent des stratégies pour gérer et co-construire des échanges, par exemple, poser des questions, être capable de prendre sa place en interaction (André, 2021) pour montrer que l'on a compris ou non.

Références bibliographiques

Adami H., André V. (2013). Corpus et apprentissage du Français Langue d'Intégration (FLI), *linx,* n°68-69, p.135-158.

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

André, V. (2021). Des corpus d'interactions dans la formation linguistique des migrants. *Savoirs, 56*(2), 77-96. https://doi.org/10.3917/savo.056.0077

André, V. (2023). De la constitution d'un corpus d'interactions à son exploitation en didactique de l'oral. Quels outils et quel accompagnement pour des pratiques innovantes ? *Recherches en didactique des langues et des cultures* [En ligne], *21*(2). https://doi.org/10.4000/rdlc.12651

Boutet, J. (2001). La part langagière du travail : bilan et évolution. *Langage et société*, 98, 17-42.

Filliettaz, L. (2009). Les discours de consignes en formation professionnelle initiale : une approche linguistique et interactionnelle. *Éducation & Didactique*, 3(1), 91-111.

Filliettaz, L. (2022). L'analyse interactionnelle comme objet et méthode de recherche. In (éd.), *Enquêter dans les métiers de l'humain : Traité de méthodologie de la recherche en Sciences de l'éducation et de la formation*. Editions Raison et Passions, 278-288.

Mondada, L. (2017). Le défi de la multimodalité en interaction. *Revue française de linguistique appliquée, XXII*(2), 71-87. https://doi.org/10.3917/rfla.222.0071

Pekarek-Doehler, S. (2006). Compétence et langage en action, *Bulletin Vals-Asla, 84*, 9-45.

Rousset, F. & Thomas, A. *(in press*). Création de ressources didactiques ciblant l'interaction pour des apprenti-es allophones en formation professionnelle : l'exemple du projet DiCoi, *Education et Socialisation.*

Veyrac, H. (2001). Aperçu de la variété des fonctions des consignes dans le monde du travail. *Pratiques,* 111/112, 77-92.

Diffusion des changements sémantiques en diachronie courte dans un corpus de tweets

Louise Tarrade [1], Jean-Philippe Magué [1] et Jean-Pierre Chevrot [2]
[1] Laboratoire ICAR (UMR5191), École Normale Supérieure de Lyon
[2] Laboratoire LIDILEM (EA 609), Université Grenoble Alpes
louise.tarrade@ens-lyon.fr, jean-philippe.magueé@ens-lyon.fr, jean-pierre.chevrot@univ-grenoble-alpes.fr

Introduction

La question des mécanismes de diffusion du changement linguistique est au cœur des préoccupations de la sociolinguistique variationniste. Des travaux majeurs ont été menés sur le sujet, notamment par Labov (1980) et Milroy (1987), qui se sont, entre autres, intéressés aux liens entre variation linguistique, circulation du changement, et réseau social. Ces études portaient le plus souvent sur le niveau phonétique-phonologique, étaient limitées géographiquement, et ne prenaient pas en compte de très grands échantillons de locuteurs. En tirant parti des corpus de médias sociaux et de la puissance de calcul disponible, la sociolinguistique computationnelle (Nguyen *et al.*, 2016) permet de dépasser ces contraintes méthodologiques inhérentes à l'époque. Les corpus de ce type permettent la reconstitution de réseaux de taille considérable associés à de grandes quantités de données textuelles. Combinées à des traitements complexes rendus possibles par la puissance des ordinateurs, ces données facilitent l'étude quantitative d'aspects moins explorés du changement linguistique. Le travail proposé est centré sur les changements sémantiques et leur diffusion, notamment en exploitant les méthodes basées sur l'hypothèse distributionnelle (Firth, 1957; Harris, 1954) comme les *embeddings* contextuels.

Nous nous intéressons ici au changement sémantique en diachronie courte et à sa diffusion au sein du réseau social formé par les utilisateurs de Twitter (maintenant nommé X). À partir d'un grand corpus de tweets dont le réseau d'utilisateurs a été reconstitué, nous définissons, pour chaque étape de la diffusion des changements, la position dans le réseau des adoptants des significations nouvelles. Nous rapprochons ensuite deux types de changements en comparant nos résultats à ceux obtenus lors de travaux précédents menés dans les mêmes conditions mais sur des innovations lexicales (Tarrade *et al.*, 2024), c'est-à-dire l'apparition et la diffusion de nouveaux mots dans le corpus. Le but est de mettre en évidence les similarités et les différences de changements affectant la forme et le sens des mots.

Corpus et méthodologie

Corpus

Le corpus sur lequel s'appuie cette recherche est composé d'environ 650 millions de tweets en français, rédigés entre 2012 et 2019[1]. Ces tweets ont été collectés en deux vagues successives.

1 https://www.ortolang.fr/market/corpora/sosweet

Le corpus initial a été construit en récupérant, par le recours à des fournisseurs de données, 170 millions de tweets produits entre 2014 et 2017 sur les fuseaux horaires GMT et GMT+1 incluant plusieurs aires francophones (Abitbol *et al.*, 2018). Au total, 2,5 millions de comptes Twitter étaient à l'origine de ces tweets. Ces comptes ont servi de base à la deuxième vague de collecte, qui s'est déroulée entre 2018 et 2019, et a permis de compléter la collecte initiale en récupérant itérativement les 3 200[2] derniers tweets de ces utilisateurs, via l'API de Twitter.

Plus précisément, l'étude de la diffusion des changements sémantiques est menée sur une période de 5 ans, correspondant aux tweets produits entre janvier 2014 et décembre 2018.

Méthodologie

Dans une première phase de nos travaux, nous avons détecté 45 formes linguistiques lieux de changements sémantiques, de différentes natures, au sein du corpus de tweets, en nous appuyant sur le déplacement de leurs vecteurs contextuels moyennés (Martinc *et al.*, 2020; Montariol *et al.*, 2021) au fil des mois. Une méthode de clustering appliquée à l'ensemble de leurs vecteurs contextuels a permis de dégager, pour chaque mot, ses principaux usages dans différents contextes. À partir de ces clusters, deux types d'usages ont été identifiés pour chaque mot : son usage initial, et son usage en expansion. Par exemple, le mot « vocaux » a subi en 5 ans un glissement catégoriel en passant d'une utilisation adjectivale (« des messages vocaux ») à une utilisation nominale (« des vocaux »). Son usage en expansion correspond à toutes ses occurrences où « vocaux » est utilisé comme nom, et ses usages initiaux à toutes les autres.

Chacun des changements sémantiques identifiés suit une trajectoire de diffusion en forme de S, typique des innovations réussies (Blythe et Croft, 2012; Rogers, 1962). En nous appuyant sur les moments où l'accélération de la diffusion a le plus varié, nous avons segmenté automatiquement ces trajectoires en trois phases (Tarrade *et al.*, 2022), qui correspondent aux trois étapes de diffusion d'un changement linguistique : une première phase d'innovation où l'usage en expansion apparaît, une seconde de propagation pendant laquelle l'usage en expansion se diffuse plus largement, et une troisième de fixation où son utilisation se maintient. Nous observons ensuite la position, dans le réseau, des individus qui utilisent les usages en expansion et les usages initiaux lors des trois étapes. Le réseau d'utilisateurs que nous avons reconstitué inclut 2.5 millions de nœuds et 330 millions de liens, qui traduisent les relations de followers/followees qu'ils entretiennent (lorsqu'un utilisateur en suit un autre ou est suivi par lui). À l'aide de la libraire Networkit (Staudt et al., 2016), permettant l'analyse de grands réseaux, nous avons caractérisé chaque utilisateur en fonction de 4 variables de réseau reflétant chacune à sa manière la position qu'il occupe au sein du réseau global ou d'une communauté particulière : prestige dans le réseau, centralité dans la communauté, enfermement dans la communauté, ouverture de son réseau égocentré.

Chaque type d'usage (initial/en expansion) de chaque mot a ensuite été caractérisé à chacune des phases de diffusion par ces 4 mesures de réseau appliquées aux utilisateurs qui les produisent. Dans le cas des usages initiaux, ces mesures caractérisent la position dans le réseau des utilisateurs qui les emploient à chaque phase. Dans le cas des usages en expansion, elles reflètent la position dans le réseau de leurs adoptants, c'est-à-dire les utilisateurs qui s'emparent de ce nouvel usage pour la première fois dans notre jeu de données. En parallèle, un groupe contrôle a été créé, composé de mots n'ayant pas manifesté de changements sémantiques, c'est-à-dire dont le vecteur contextuel moyenné ne s'est pas déplacé au fil du temps. Ces mots ont

2 Limite imposée par l'API Twitter.

été caractérisés de la même manière en fonction des mesures de réseau des utilisateurs qui les emploient. Au final, des tests univariés ont permis de comparer la distribution des variables de réseau entre ces différents groupes et à différentes périodes : à chacune des phases de diffusion, mais aussi dans la période précédant le changement. Nous mettons également en regard la dynamique de diffusion des changements sémantiques et celle de changements lexicaux.

Résultats

L'analyse de la diffusion à partir des variables de réseau caractérisant les usages en expansion, les usages initiaux, et les mots du groupe contrôle, ont révélé un profil comparable pour les deux types d'usage. Néanmoins, leur distinction avec la population du groupe contrôle et les similitudes avec les résultats obtenus précédemment sur les changements lexicaux (Tarrade *et al.*, 2024) nous ont poussé à nous intéresser à la période précédant le changement, c'est-à-dire avant que le changement sémantique n'ait été identifié. Les résultats obtenus ont permis de montrer que les mots en passe d'être l'objet d'un changement sémantique se distinguent déjà des mots du groupe contrôle en étant utilisés par des individus au réseau plus ouvert, qui sont moins prestigieux dans le réseau global et moins centraux dans leur communauté.

Ces mêmes caractéristiques ont été observées lors de la première phase de diffusion d'innovations lexicales, et la mise en regard de ces deux types de changement nous permet de constater qu'ils suivent également une dynamique de diffusion similaire, comme le montre la figure 1.

		Innovation → Propagation	Propagation → Fixation
Ouverture du réseau	Usages en expansion	↗	
	Changements lexicaux	↗	↗
Prestige	Usages en expansion	↗	↗
	Changements lexicaux	↗	↗
Centralité	Usages en expansion	↗	↗
	Changements lexicaux	↗	↗
Enfermement dans la communauté	Usages en expansion	↘	↘
	Changements lexicaux		↘

figure . 1 Dynamique des variables de réseau entre les différentes phases de diffusion des usages en expansion et des changements lexicaux

Par exemple, les adoptants d'usages en expansion, comme ceux des changements lexicaux, ont un réseau égocentré de plus en plus ouvert (ce qu'indique la flèche montante) entre la phase d'innovation et de propagation. En revanche, la différence observée pour cette même variable entre la phase de propagation et de fixation n'est pas significative pour les usages en expansion (comme l'indique le fond hachuré). Le prestige et la centralité des adoptants des deux types de changement tendent à augmenter au fil des phases, et ils sont de moins en moins enfermés dans leur communauté. Si la dynamique de diffusion est très similaire, nous observons cependant des valeurs de centralité et de prestige plus hautes pour les adoptants de changements sémantiques. Nous expliquons ce phénomène par le processus moins conscient d'acquisition pour les changements sémantiques, plus sensibles à un apprentissage implicite par répétition. Il s'avère en effet que ces variables de réseau sont des mesures très liées à la fréquence d'exposition aux usages linguistiques.

Nous concluons de ces résultats que les individus dont la position dans le réseau est marquée par une centralité et un prestige faible, ainsi qu'un réseau égocentré assez ouvert sont plus enclins à l'innovation. Au fil des phases, les changements sont adoptés par des individus de plus en plus centraux et prestigieux, conservant un réseau égocentré ouvert et peu enfermés

dans leur communauté. Ces observations rejoignent, précisent et concilient les différents profils décrits dans les travaux antérieurs menés en sociolinguistique variationniste (Labov, 1980; Milroy, 1987), malgré des différences notables d'échelle, de méthodologie, et de types de changements étudiés (lexicaux, sémantiques, phonétiques). Elles permettent également d'obtenir une vue d'ensemble de la diffusion des changements sémantiques dans un réseau social en diachronie courte, aspect encore peu étudié à notre connaissance dans les travaux computationnels sur les changements sémantiques.

Références bibliographiques

Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P. et Fleury, E. (2018). Socioeconomic Dependencies of Linguistic Patterns in Twitter: a Multivariate Analysis. Dans *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18* (p. 1125-1134). ACM Press. https://doi.org/10.1145/3178876.3186011

Blythe, R. A. et Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, *88*(2), 269-304. JSTOR.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 10-32.

Harris, Z. S. (1954). Distributional Structure. *Word*, *10*(2-3), 146-162. https://doi.org/10.1080/00437956.1954.11659520

Interactions, corpus, apprentissages et représentations – UMR 5191 (ICAR), DANTE Inria, Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles – EA 609 (LIDILEM), et Modélisation et analyse linguistique automatique et humanités computationnelles (ALMANACH). (2024). *SoSweet* (version v1) [corpus]. ORTOLANG (Open Resources and TOols for LANGuage). https://hdl.handle.net/11403/sosweet/v1

Labov, W. (1980). The social origins of sound change. Dans *Locating Language in Time and Space* (p. 251-265). Academic Press New York.

Martinc, M., Novak, P. K. et Pollak, S. (2020). Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift. Dans *Proceedings of the 12th Language Resources and Evaluation Conference* (p. 4811-4819).

Milroy, L. (1987). *Language and social networks* (2nd ed). B. Blackwell.

Montariol, S., Doucet, A. et Allauzen, A. (2021). État de l'art du changement sémantique à partir de plongements contextualisés. Dans *COnférence en Recherche d'Informations et Applications-CORIA 2021, French Information Retrieval Conference*.

Nguyen, D., Doğruöz, A. S., Rosé, C. P. et de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational linguistics*, *42*(3), 537-593. https://doi.org/10.1162/COLI_a_00258

Rogers, E. M. (1962). *Diffusion of innovations*. Free Press.

Staudt, C. L., Sazonovs, A., & Meyerhenke, H. (2016). NetworKit: A tool suite for large-scale complex network analysis. Network Science, 4(4), 508-530. https://doi.org/10.1017/nws.2016.20

Tarrade, L., Chevrot, J.-P. et Magué, J.-P. (2024). How position in the network determines the fate of lexical innovations on Twitter. *PLOS Complex Systems*, *1*(1), e0000005. https://doi.org/10.1371/journal.pcsy.0000005

Tarrade, L., Magué, J.-P. et 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

Dire le corps en mouvement : noms de parties du corps et consignes verbales dans la danse contemporaine et du fitness fonctionnel

Chiara Minoccheri [1], Angelina Aleksandrova[2]
[1] Laboratoire CLLE, Université Toulouse Jean Jaurès
[2] Laboratoire EDA 4071, Université Paris Cité
chiara.minoccheri@univ-tlse2.fr ; angelina.aleksandrova@u-paris.fr

Introduction

Cette contribution s'inscrit dans une recherche plus large consacrée à l'étude des discours procéduraux produits en contexte d'activité physique (projet FITTALK[1]), avec une focale particulière sur les consignes verbales en danse contemporaine et le fitness fonctionnel (séances de CrossFit®[2]). Dans cette étape du projet, nous nous intéressons à l'emploi des noms de parties du corps (NPC) dans ces deux disciplines, dans l'objectif d'examiner la manière dont le corps est segmenté, référé et mis en mouvement dans le langage des instructeurs. Deux questions guident notre réflexion : (i) quelles parties du corps sont mobilisées dans les consignes et (ii) que révèlent ces formes d'encodage sur les représentations conceptuelles du corps en action ?

Cette étude repose sur l'hypothèse que la nature contrastée des deux disciplines – l'une tournée vers l'expressivité artistique et la création gestuelle, l'autre vers la performance physique ritualisée – se reflète dans les formes linguistiques qu'elles mobilisent pour référer au corps. En danse contemporaine, les mouvements sont rarement associés à des dénominations techniques stables. Hormis quelques emprunts à la danse classique ou aux sciences du corps (ex. spirale, suspension, pont), les consignes font appel à un lexique courant, parfois métaphorique, et varient selon les chorégraphes (Minoccheri, 2023; Minoccheri et al., 2024 ; Minoccheri & Aurnague, 2021). À l'inverse, les séances de crossfit sont fortement ritualisées et les exercices référés par des désignations plutôt stabilisées – souvent en anglais ou sous forme polylexicale (Aleksandrova 2023, Aleksandrova 2025 à par.). Alors que la danse contemporaine est une activité physique visant, avant tout, l'expression artistique au travers du mouvement corporel, le crossfit désigne, dans le langage courant, une méthode d'entraînements croisés visant la polyvalence athlétique en développant aussi bien l'endurance cardiovasculaire, pulmonaire et musculaire que la force, la souplesse, la puissance, la vitesse, l'équilibre, etc. (cf. https://www.crossfit.com). Cette différence d'approche affecte les usages linguistiques : là où la danse verbalise l'exploration gestuelle, le crossfit actualise un répertoire de mouvements attendus.

1 Le projet FITTALK a bénéficié d'un soutien dans le cadre du programme Investissement d'Avenir - Emergence, mis en œuvre par l'ANR (ANR-18-IdEx-0001).

2 A l'origine il s'agit d'une entreprise américaine, leader dans l'industrie privée du sport au niveau mondial. Dans l'usage, l'appellation crossfit désigne la pratique d'entrainements croisés de fitness fonctionnel.

À long terme, l'intérêt de ce travail est double. Il s'agit d'abord de mettre en regard deux formes d'encadrement verbal du corps, qui malgré leurs différences, partagent un même ancrage fondamental : l'attention portée au geste, à sa précision, à sa coordination. Les consignes orales, qu'elles soient artistiques ou sportives, constituent des objets linguistiques riches, où s'articulent description, injonction, guidance et co-construction du mouvement. Ensuite, alors qu'un bon nombre de travaux portant sur diverses activités physiques se situent dans les champs de la linguistique interactionnelle et de l'analyse conversationnelle et s'intéressent principalement aux aspects multimodaux de la communication (voir par exemple Albert et vom Lehn 2023; Broth et Keevallik 2014; Keevallik 2010, 2013, 2021; Råman et Haddington 2018), notre recherche entend combler un angle mort des travaux en sémantique sur les discours en contexte d'apprentissage corporel. En apportant une contribution originale à la linguistique de corpus, elle documente des formats discursifs peu explorés, à la fois situés, oraux, et intimement liés à la dynamique corporelle. Cela étant dit, notre objectif principal ici sera de comparer les deux sources de données à l'échelle de l'emploi des noms de parties du corps, première étape vers une caractérisation linguistique plus complète des consignes de mouvement, à l'intersection entre activité corporelle, verbalisation de l'effort, et gestion pédagogique du geste. Elle entend apporter un éclairage inédit sur la manière dont les discours d'encadrement – qu'ils soient artistiques ou athlétiques – donnent forme au corps, en conjuguant injonction, description, et modulation attentionnelle.

Corpus et méthodologie

Cette étude repose sur deux corpus oraux transcrits, contrastés quant à leur ancrage disciplinaire, leur ritualisation et leur visée, mais comparables du point de vue du genre discursif des consignes verbales adressées à des pratiquants en contexte d'activité corporelle.

Corpus

Le corpus CominDanse[3], constitué dans le cadre de la thèse de Minoccheri (2023), rassemble quatre leçons de danse contemporaine de niveau avancé à professionnel, données par quatre chorégraphes-pédagogues français (deux hommes et deux femmes de quatre générations différentes). Ce niveau est défini par une pratique quotidienne au sein d'une compagnie ou d'un cursus supérieur en danse. Les séances, totalisant 7h16, ont été intégralement transcrites et segmentées en énoncés à l'aide du logiciel ELAN (Wittenburg et al., 2006), selon des critères syntaxiques et pragmatiques (Gary-Prieur, 1985 ; Muller, 2002). Seuls les énoncés directifs orientés vers l'exécution du mouvement corporel ont été retenus (Austin, 1962 ; Searle, 1972 ; Moneglia & Raso, 2014), pour un total de 35 916 mots. Le corpus COWOD[4], constitué entre 2020 et 2022 par A. Aleksandrova, est issu de 105 vidéos de séances de CrossFit® diffusées en ligne pendant les confinements liés à la pandémie de Covid-19. Ces capsules, d'une durée moyenne de trois minutes, correspondent à la présentation de la séance du jour, moment clé durant lequel le coach expose les objectifs de l'entraînement, explicite les mouvements, propose des adaptations et motive les participants. Les vidéos ont été produites par dix coachs certifiés, tous de genre masculin, et visent à maintenir la dynamique collective à distance. Le corpus totalise 85 000 mots ; un sous-échantillon de 36 000 mots a été extrait afin d'équilibrer la comparaison avec le corpus CominDanse. Il est important de préciser que ces vidéos se distinguent par une situation énonciative asymétrique et décontextualisée par rapport aux cours de danse ou des séances d'entraînement crossfit *in situ*. Cependant, bien qu'il s'agisse d'une

3 COrpus Multimodal d'INstructions de Danse contemporaine.

4 Mot valise entre covid et WOD (workout of the day, angl. « entraînement du jour »).

énonciation différée, ces enregistrements se rapprochent assez de ce qui se passe lors d'un entraînement dans la salle parce qu'ils correspondent à un moment-clé – l'introduction – de chaque wod, conformément à la méthodologie crossfit :

> Bien qu'elle soit la partie la moins longue du cours, l'introduction est très importante. Pendant l'introduction (qui se déroule généralement au tableau blanc), l'entraîneur présente l'entraînement, définit les attentes associées en expliquant le stimulus visé, propose différentes options d'adaptation et répond aux éventuelles questions des participants. En général, l'introduction prend deux à quatre minutes, selon la complexité de l'entraînement et le nombre de participants au cours. (Crossfit Inc. (2020), Guide d'entrainement et de révision - niveau 2, p. 67).

Méthodologie

Notre démarche croise des approches quantitatives et qualitatives. L'analyse textométrique, menée sur les corpus lemmatisés, a permis d'extraire les énoncés contenant un nom de partie du corps (NPC), à partir d'une liste de lemmes unifiée. Les NPC les plus fréquents ont été isolés, puis analysés en contexte. Les analyses qualitatives, menées sur les directives impliquant les NPC les plus fréquents, consisteront à étudier plusieurs facteurs dans l'expression du mouvement des parties du corps mises en jeu. Premièrement, le rôle sémantique endossé par les parties du corps au sein de la description spatiale que l'énoncé sous-tend. Les NPC peuvent désigner des parties du corps localisées ou mises en mouvement (le bassin est poussé légèrement vers l'avant - COWOD) et ils endossent alors le rôle de la cible. Si, en revanche, la partie du corps (ici le coccyx) sert de repère pour le mouvement d'une autre entité (les ischions basculent en direction du coccyx - CominDanse), on la qualifie de site (Vandeloise, 1986). Deuxièmement, nous étudierons la nature des cooccurrents indiquant les mouvements dans lesquels les parties du corps sont impliquées.

Résultats

Les résultats préliminaires de l'analyse quantitative révèlent une distribution contrastée des noms de parties du corps dans les deux corpus. Dans CominDanse, les dix items les plus fréquents sont : *pied, jambe, colonne, tête, bras, bassin, main, cage, ischion, talon*. Dans COWOD, on trouve : *bras, jambe, poitrine, dos, tête, pied, bassin, genou, main, épaule*. Cette variation suggère d'emblée une appréhension différenciée du corps dans les deux pratiques. La danse contemporaine met en avant une représentation centrée sur la structure interne et l'ancrage osseux du corps (colonne vertébrale, cage thoracique, ischion), témoignant d'une attention à l'alignement et aux axes de mouvement. À l'inverse, le discours des coachs de CrossFit privilégie une vision plus fonctionnelle et segmentée, fondée sur des repères musculaires ou extérieurs (poitrine, dos, épaule), en lien avec l'efficacité gestuelle et la performance. Afin d'établir une base comparative robuste, notre analyse qualitative se concentre sur les six NPC les plus fréquemment partagés entre les deux corpus : *bras, jambe, pied, tête, bassin* et *main*. L'étude des cooccurrents verbaux associés à ces items met en lumière des régularités contrastées. Par exemple, les cooccurrents de *bras* avec l'indice de fréquence le plus élevé sont *tendre* dans COWOD, renvoyant à une injonction technique, et *accompagner* dans CominDanse qui traduit plutôt une logique d'orientation douce ou de guidance du geste. De même, *pied* est associé à *toucher* dans le CrossFit (recherche de contact ou de point d'appui), alors qu'en danse, il cooccurre avec *lancer*, témoignant d'une dynamique d'expansion spatiale.

Ces observations, qui seront développées dans notre présentation, soulignent l'intérêt de confronter des corpus issus de pratiques corporelles contrastées pour mettre au jour des régularités langagières indexées sur des régimes d'action distincts. Elles permettent également

d'identifier des configurations syntaxico-sémantiques spécifiques aux consignes de mouvement dans des environnements pédagogiques oraux, entre chorégraphe-pédagogue et danseurs d'une part, et entre coach et pratiquants d'autre part. Dans un champ encore peu exploré par la linguistique de corpus, cette étude entend ainsi apporter des éclairages méthodologiques et théoriques sur les discours du corps en action, au croisement du langage, du geste, de la cognition et de l'apprentissage situé.

Références bibliographiques

Albert, S., & vom Lehn, D. (2023). Non-lexical vocalizations help novices learn joint embodied actions. *Language & Communication*, *88*, 1-13. https://doi.org/10.1016/j.langcom.2022.10.001

Aleksandrova, A. (2023). Catégorisation et vocabulaire spécialisé : Enquête sur les dénominations de mouvements sportifs. Lexis. Journal in English Lexicology, 22. https://journals.openedition.org/lexis/7366

Aleksandrova, A. (2025, à paraître), « Dénominations polylexicales et polysémie : l'exemple des mouvements de la préparation physique », in L'Homme M.-Cl. & Frasso P. numéro thématique *En termes de polysémie : sens et polysémie dans les domaines de spécialité*, Repères -DORIF.

Austin, J. L. (1962). How to Do Things with Words. Clarendon Press.

Broth, M., & Keevallik, L. (2014). Getting Ready to Move as a Couple : Accomplishing Mobile Formations in a Dance Class. *Space and Culture*, *17*(2), 107-121. https://doi.org/10.1177/1206331213508483

Gary-Prieur, M.-N. (1985). De la grammaire à la linguistique : L'étude de la phrase. Armand Colin.

Keevallik, L. (2010). Bodily Quoting in Dance Correction. Research on Language and Social Interaction, 43(4), 401-426. https://doi.org/10.1080/08351813.2010.518065

Keevallik, L. (2013). Here in time and space : Decomposing movement in dance instruction. In P. Haddington, L. Mondada, & M. Nevile (Éds.), Interaction and Mobility : Language and the Body in Motion. De Gruyter.

Keevallik, L. (2021). Vocalizations in dance classes teach body knowledge. Linguistics Vanguard, 7(s4). https://doi.org/10.1515/lingvan-2020-0098Minoccheri, C. (2023). L'espace dans le corps, le corps dans l'espace : L'expression linguistique du mouvement en danse contemporaine [PhD Thesis]. Université Toulouse Jean Jaurès.

Minoccheri, C., & Aurnague, M. (2021). Agir sur le corps pour se mouvoir dans l'espace. Étude sémantique des verbes causatifs de mouvement à partir d'un corpus d'instructions de danse. Lexique, 28, 63-86.

Minoccheri, C., Combe, C., & Stosic, D. (2024). À la manière de la presse écrite : L'expression de la manière dans un corpus journalistique en comparaison avec deux autres genres discursifs. SHS Web of Conferences, 191, 12003. https://hal.science/hal-04696669/

Moneglia, M., & Raso, T. (2014). Notes on the Language into Act Theory. In T. Raso & H. Mello (Éds.), Spoken Corpora and Linguistic Studies (p. 468-495).

Muller, C. (2002). Les bases de la syntaxe : Syntaxe contrastive, français—Langues voisines. Presses universitaires de Bordeaux.Råman, J., & Haddington, P. (2018). Demonstrations in Sports Training : Communicating a Technique through Parsing and the Return-Practice in the Budo Class. Multimodal Communication, 7(2). https://doi.org/10.1515/mc-2018-0001

Searle, J. R. (1972). Les actes de langage : Essai de philosophie du langage (H. Pauchard, Trad.). Hermann.

Vandeloise, C. (1986). L'espace en français : Sémantique des prépositions spatiales. Éditions du Seuil.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. Proceedings of the 5th International Conference on Language Resources and Evaluation, 1556-1559.

Discourse Markers in Contact : Investigating *'Like'* and *'Parang'* in Tagalog-English Bilingual Speech

Patricia Camus[1] and Lisa Brunetti[1]
[1] Laboratoire de Linguistique Formelle (LLF – UMR7110), Université Paris Cité/ CNRS
patricia.camus@u-paris.fr, lisa.brunetti@u-paris.fr

Introduction

Discourse markers (DMs) have been broadly defined as a class of linguistic items that: (1) contribute to discourse coherence (Schiffrin, 1987); (2) function as signposts that indicate the relevance of one discourse segment to another (Blakemore, 1987); and (3) help speakers express their evaluation, attitudes, and beliefs towards the propositional content of an utterance (Fraser, 1996), among many other definitions.

In bilingual speech, DMs are particularly interesting due to their high susceptibility to contact-induced changes (Matras, 2007). Tagalog, an Austronesian language spoken in the Philippines, provides an interesting case in the study of bilingual DMs in language contact situations due to its extensive and long-standing contact with English. As both Tagalog and English hold official status in the country, majority of Filipinos are exposed to both languages from an early age, resulting to frequent code-switching especially in informal settings.

This study examined a pair of bilingual DMs, namely the English *like* and the Tagalog *parang* /ˈpaɾaŋ/, as they are used in Tagalog-English speech. While there is a substantial body of research on *like* in Inner Circle Englishes (Buchstaller, 2001; Fuller, 2003) and in Philippine English (Schweinberger, 2014), relatively little attention has been given to *like* in Taglish speech. Similarly, *parang* has been examined within monolingual Tagalog contexts (Schachter & Otanes, 1972; Malicsi, 2013), but its role in bilingual discourse remains largely unexplored. This study sought to fill this gap by analyzing *like* and *parang* in bilingual discourse, focusing on their clausal position, discourse functions, and language context.

Corpus and Methodology

A corpus-based approach was used to investigate *like* and *parang*. Fifteen Filipino bilinguals (n = 15; 8 females, 7 males; M age = 28.5 years, range = 19–43) were interviewed online. Each participant was asked to respond to five personal questions, besides a final task of retelling the Pear Story (Chafe, 1980). All interviews were audio-recorded using kMeet, fully transcribed in Praat (Boersma & Weenik, 2023), and manually segmented into clauses based on Nagaya's (2007) work on the layered structure of the clause in Tagalog. The resulting interlinear TextGrids were then converted into .csv files for easier manipulation in AntConc (Anthony, 2023). The total audio material amounted to around eight hours, and yielded a spoken corpus of 78,297 tokens.

For the annotation of position, the segmentation into clauses enabled the identification of positions in which the DMs may occur, namely, Clause-Initial, Clause-Medial, and Clause-Final (although Clause-Final *like* and *parang* have not been attested in the current corpus).

For the functional annotation of *like* and *parang*, previous studies were reviewed which proposed various terminologies for similar or overlapping functions (Buchstaller, 2001; Schweinberger, 2014, a.o. for *like* ; Schachter & Otanes, 1972 ; Tan, 2024, a.o. for *parang*). These previously proposed functions were synthesized and consolidated to come up with a list of functions for *like* and *parang*, while remaining attentive to the possibility that certain functions might not fully align across the two markers, or that the data may be indicating other functions not on the list. The functions can be grouped into two broad categories: DM functions and Non-DM functions. DM uses include Elaboration (*For long term, imposible talaga kasi ang mahal niya. LIKE one thousand two hundred yung binabayad ko before*. 'For long term, it's really impossible because it (the apartment) is really expensive. LIKE I was paying one thousand two hundred before.'), Hedging (*Compared sa Pilipinas, dito kasi sa Europe, PARANG very individualized ang kanilang lifestyle.* 'Compared to the Philippines, here in Europe, IT SEEMS LIKE their lifestyle is very individualized.'), and Hesitation (*Sa simula they had LIKE uhm uhm a little bit of a stipend.* 'At the beginning they had LIKE uhm uhm a little bit of a stipend.'), while Non-DM uses include Marking Focus, Marking Quotations, and Marking Comparisons.

Lastly, we also looked at the language context of each instance of *like* and *parang*, noting whether adjacent words were in English or Tagalog. Four distinct environments were identified: TAG-TAG, TAG-ENG, ENG-TAG, and ENG-ENG.

Results

A generalized linear model (GLM) was used to examine the factors that predict the use of *like* and *parang* in the corpus. The predictors included clausal position (Clause-Initial, Clause-Medial), discourse function (Elaboration, Hedging, Hesitation, Non-DM Use), and language context (TAG-TAG, TAG-ENG, ENG-TAG, ENG-ENG). The corpus contained a total of 223 *like* occurrences and 710 *parang* occurrences. Of these figures, a stratified random sample of 170 tokens each of *like* and *parang* have been annotated for the present study.

When it comes to clausal position, results show that there is no significant difference between *like* and *parang*. Both DMs tend to occur clause-initially. Functionally, it appears that *like* occurs significantly more likely than *parang* in Elaboration contexts and significantly less likely to occur with a Hedging function ($p < .001$), also significantly less likely to occur with a Non-DM function ($p < .001$), and only marginally less often with a Hesitation function ($p < .067$). Lastly, when it comes to language context, *like* is significantly more likely to occur in fully English utterances (ENG-ENG), but less likely to appear in mixed or fully Tagalog utterances. More precisely, *like* is less likely than *parang* in ENG-TAG, TAG-ENG, and TAG-TAG (all of which are $p < .001$) language contexts.

Discussion and Conclusion

The current results show the co-existence of *like* and *parang* in Tagalog–English speech. Despite both markers manifesting all DM and non-DM functions, speakers do not exploit *like* to the same extent as *parang*. *Like* is used primarily for Elaboration, while *parang* is used primarily for hedging and non-DM uses. What this signals might be the emergence of a division of labor between the two markers: *like* is becoming pragmatically specialized for Elaboration, while *parang* is maintaining a wider functional range. Additionally, the significant drop in the likelihood of *like* in Tagalog or mixed discourse suggests that it currently functions as a code-switched element rather than a fully integrated borrowing. While eventual borrowing remains possible, the present results suggest that *like* still remains dependent on English in the discourse.

Taken together, these findings suggest that *like* and *parang* may be developing along different grammaticalization paths: *like* remains constrained and pragmatically specialized in English contexts, while *parang* undergoes semantic and functional expansion within Tagalog discourse. Traditionally described as a comparative/similative marker (Santos, 1939; Schachter & Otanes, 1972), *parang* now also operates as a DM indicating hedge as well as a non-DM marking focus and introducing quotations. Although English enjoys greater prestige in Philippine society, it is Tagalog that continues to be used more widely in everyday interaction. Perhaps this broader social embeddedness allows *parang* to expand and diversify its functions, while *like* remains more restricted to English-dominant contexts. From a grammaticalization perspective, this suggests that frequency of use and depth of integration in everyday discourse are crucial drivers of functional expansion, enabling *parang* to evolve while constraining *like*'s trajectory.

Bibliographic References

Anthony, L. (2023). AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Blakemore, D. (1987). Semantic constraints on relevance. Oxford: Blackwell.

Boersma, P., & Weenink, D. (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.18, retrieved 8 October 2023 from http://www.praat.org/

Chafe, W. 1980. (ed.), The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production. Norwood, New Jersey: Ablex.

Fuller, J. (2003). Use of the discourse marker like in interviews. *Journal of Sociolinguistics - J SOCIOLING.* 7. 365-377. 10.1111/1467-9481.00229.

Fraser, B. (1996). Pragmatic markers. *Pragmatics*, 6 (2), 167-190.

Malicsi, J. (2013). Gramar ng Filipino. Quezon City: Sentro ng Wikang Filipino.

Matras, Y. (2007). The borrowability of grammatical categories. In Yaron Matras & Jeanette Sakel (eds.), Grammatical borrowing in cross-linguistic perspective. Berlin: Mouton De Gruyter, 31–74.

Miller, J., & Weinert, R. (1995). The function of LIKE in dialogue. *Journal of Pragmatics*, 23, 365-393.

Nagaya, N. (2007). Information Structure and Constituent Order in Tagalog. *Language and Linguistics*. 1. 343-372.

Santos, L. (1939). Balarila ng Wikang Pambansa. P.I. Surian ng Wikang Pambansa. Bureau of Printing.

Schachter, P. & Otanes, F. T. (1972). Tagalog reference grammar. Berkeley: University of California Press.

Schiffrin, D. (1987). Discourse markers. Cambridge/New York/Melbourne: Cambridge University Press.

Schweinberger, M. (2014). The discourse marker LIKE: a corpus-based analysis of selected varieties of English. 10.13140/RG.2.2.28150.65603.

Tan, B. (2024). A study on the pragmatic functions of parang in Tagalog utterances. *The Archive*. Regular Issue Vol. 5, No. 1. Available at: https://journals.upd.edu.ph/index.php/archive/article/view/10054 (Accessed 24 April 2025)

Entre référence, thématisation et création référentielle : vers une typologie discursive des mots en *qu-* en français

Axelle Domingues[1]
[1] Laboratoire STIH, Sorbonne Université
axelle.domingues@sorbonne-universite.fr

Introduction

Les mots en *qu-* constituent à première vue un paradigme unifié. D'un point de vue formel, ils prennent l'initiale commune en qu- venant des racines indo-européennes *k^we/o- et *k^wei- (Meillet & Vendryès, 1968 : §§ 750-752 ; Szemerényi, 1999 : 208). D'un point de vue fonctionnel, ils partagent la notion de variable transmise par ces racines qui exprimaient au départ simultanément l'indéfini et l'interrogatif (Le Goffic, 2002, 2007 : 20). Au cours des siècles, ce paradigme unique s'est fragmenté en trois sous-paradigmes distincts : les indéfinis, les interrogatifs et les relatifs. Cette spécialisation fonctionnelle représente la catégorisation traditionnelle des mots en *qu-*. Cependant, l'admettre sans remise en question reviendrait à ignorer la complexité posée par ce paradigme brisé. Les différents sous-paradigmes présentent eux-mêmes des failles caractéristiques de la porosité des frontières qui les séparent. Cet aspect révèle de nombreuses irrégularités dans la séparation traditionnelle des mots en *qu-* qui appelle à des considérations plus subtiles sur cet objet d'étude.

La complexité du traitement de la configuration de la référence et des fonctionnements discursifs illustre de façon parlante la perméabilité des sous-paradigmes des mots en *qu-*. Notre recherche vise à proposer une typologie discursive permettant de reconsidérer cette fragmentation. Par *configuration de référence*, nous entendons les modalités par lesquelles une forme en *qu-* introduit, relaye ou construit un référent dans le discours. Par *fonctionnements discursifs*, nous désignons leur rôle dans la cohésion textuelle, à travers la reprise, la thématisation ou la création de référents. L'objectif est de déterminer si ces fonctionnements respectent les frontières définies par la grammaire traditionnelle, ou s'ils révèlent la persistance de continuités héritées de leur origine commune.

Corpus et méthodologie

Corpus

L'étude s'appuie sur trois corpus couvrant l'évolution du français du Ve siècle à nos jours : PaLaFra-Lat (latin tardif), la Base de Français Médiéval (IX^e-XV^e) et Frantext (X^e siècle au français contemporain). Les occurrences des formes en *qu-* à valeur relative, interrogative et indéfinie ont été extraites via le logiciel TXM selon des requêtes lemmatisées en CQL (*Corpus Query Language*).

Méthodologie

L'analyse se déploie en deux étapes complémentaires. La première est qualitative : elle consiste à élaborer une typologie des chaînes référentielles à partir des occurrences en *qu-*. Trois

configurations sont distinguées : (1) les chaînes explicites, où la forme en *qu-* reprend un antécédent lexicalisé (par exemple *La femme qui parle*) ; (2) les chaînes implicites, où des périphrases comme *ce que, ce dont, ce qui* introduisent un contenu propositionnel ou abstrait ; (3) les chaînes non référentielles, où la forme en *qu-* construit un référent générique ou hypothétique, sans ancrage lexicalisé (comme dans *Qui m'aime me suive*). Cette typologie repose sur une analyse croisée de la structure syntaxique (présence ou absence d'antécédent, distinction entre sujet et objet), du statut sémantique du référent (animé ou inanimé) et de la fonction discursive (reprise, thématisation, création).

La seconde étape est quantitative. Elle s'appuie directement sur l'annotation manuelle produite lors de l'analyse qualitative, qui permet de coder chaque occurrence selon la typologie retenue. Compte tenu du volume considérable des données, l'étude adoptera un principe d'échantillonnage équilibré par siècle et par genre textuel, de manière à assurer une représentativité diachronique et stylistique sans viser l'exhaustivité. À partir de cet échantillon, les fréquences relatives des différents types de chaînes seront calculées et croisées avec les fonctions syntaxiques et discursives des formes en *qu-*. Enfin, des scripts Python (Pandas, Numpy, Matplotlib, Seaborn) permettront de produire des visualisations destinées à mettre en évidence les grandes tendances diachroniques et les zones de porosité entre catégories.

Résultats

L'application de la typologie proposée aux données extraites devrait permettre de dégager des régularités significatives dans la répartition diachronique des trois types de chaînes identifiées. On s'attend à ce que les chaînes référentielles explicites dominent dans les premières phases de l'histoire du français, en lien avec la stabilité syntaxique du paradigme relatif à antécédent. Les chaînes référentielles implicites, fondées sur les périphrases en *ce que*, *ce dont*, etc., devraient apparaître plus tardivement, notamment à partir du moyen français, et connaître une expansion notable dans les usages discursifs structurés du français classique. Enfin, les chaînes non référentielles, qui construisent un référent non spécifié ou générique sans ancrage lexicalisé, devraient se maintenir en usage marginal ou spécialisé, tout en montrant des dynamiques intéressantes dans certains genres (juridique, religieux, maximes, etc.).

Cette typologie permettra de réévaluer la cohérence interne et les frontières fonctionnelles des sous-paradigmes en *qu-*. Elle devrait notamment confirmer la porosité structurelle entre les formes relatives sans antécédent, les interrogatives directes et certaines indéfinies, toutes caractérisées par l'absence d'ancrage référentiel fort. À l'inverse, les formes périphrastiques en *ce-* confirmeraient leur ancrage dans une structuration discursive orientée vers la thématisation de contenus abstraits, probablement lié à la présence du pronom démonstratif *ce* dans la forme.

Références bibliographiques

Le Goffic P. (2002). « Marqueurs d'interrogation / indéfinition / subordination : essai de vue d'ensemble », *Verbum* XXI(4), 315-340.

Le Goffic, P. (2007). Les mots *qu-* entre interrogation, indéfinition et subordination : quelques repères. *Lexique 18 : Les mots en qu- du français*, 13-46.

Meillet, A. & J. Vendryes (1968). *Traité de grammaire comparée des langues classiques*. Paris : Champion.

Szemerényi, O. (1999). *Introduction to Indo-European Linguistics*. Oxford: Oxford University Press.

Étude de l'accord dans des textes annotés sans consignes précises : le cas de la charge mentale dans des témoignages professionnels

Iuliia Arsenteva [1,2], Caroline Dubois [1], Philippe Le Goff [1], Sylvie Plantin [1] et Ludovic Tanguy [2]
[1] Orange Innovation [2] Laboratoire CLLE, CNRS - Université Toulouse 2
Arsenteva Iuliia (iuliia.arsenteva@orange.com), Dubois Caroline (caroline.dubois@orange.com), Le Goff Philippe (philippe2.legoff@orange.com), Plantin Sylvie (sylvie.plantin@orange.com), Tanguy Ludovic (ludovic.tanguy@univ-tlse2.fr)

Introduction

Les campagnes d'annotation linguistique de texte visent généralement l'application d'une grille spécifique en visant un consensus, même si plusieurs annotateurs participent à l'effort, en retenant au final les propositions majoritaires (Pustejovsky et al., 2017). De plus en plus, la communauté du Traitement Automatique du Langage remet en question cette méthode, suggérant de considérer la multiplicité des annotations, ou plus précisément la diversité des perspectives dans l'annotation, en les intégrant aux modèles lors de l'apprentissage. Cette approche, qui vise essentiellement les phénomènes sémantiques avec une forte subjectivité (opinion, sentiment, pragmatique...) porte le nom de *perspectivisme* et a été formulée par Cabitza et al. (2023). Le sujet soulève un ensemble de questions à la fois méthodologiques (Plank, 2022) et éthiques (Valette, 2024).

Malgré un intérêt croissant du public pour la notion de la charge mentale, il n'existe pas de définition consensuelle (Cain, 2007). Mais ce concept est très important dans le contexte de l'entreprise, car il est lié au bien-être des salariés. L'étude de ce concept peut contribuer à la qualité de vie au travail. Nous considérons ici que la notion de charge mentale est l'effort mental permettant d'anticiper et d'investir les ressources nécessaires, en fonction de la quantité disponible, pour la réalisation de la composante mentale de la tâche (Le Gonidec, 2022).

Ce travail s'inscrit dans un projet de recherche et développement dans une grande entreprise, qui vise à recueillir des messages issus de l'expression de salariés, et à partir de ceux-ci d'extraire des indicateurs psycho-sociaux, marqueurs de la relation aux autres, à l'entreprise, et notamment de la charge mentale.

Nous avons appliqué le principe du perspectivisme en le poussant un peu plus loin en utilisant une tâche d'annotation pour explorer les différentes façons dont est perçue et intégrée la notion de charge mentale dans ces messages écrits, par des utilisateurs (eux-mêmes salariés de l'entreprise) n'ayant pas d'expérience de l'annotation.

Plus techniquement, il s'agit ici de comparer les résultats des annotations d'un petit ensemble de textes et d'identifier les points de convergence et de différence. Ceci a aussi pour but de voir s'il existe pour ces utilisateurs un "socle" commun de connaissances sur le thème de la charge mentale, et d'en identifier des indicateurs.

Corpus et méthodologie

Corpus

Le corpus utilisé dans cette étude est constitué de données recueillies à l'aide d'un outil interne développé dans l'entreprise, qui permet aux salariés de partager leurs expériences, leurs idées et leurs pensées de façon spontanée et anonyme, autour d'une thématique donnée.

Pour cette étude, nous avons choisi 8 messages (entre 78 et 245 mots) issus de diverses campagnes portant sur des sujets tels que le retour d'expérience suite à un déménagement vers un nouveau site, ou un événement majeur de la vie de l'entreprise. Aucun des verbatims ne parle explicitement de la charge mentale de travail.

Méthodologie

Pour notre expérimentation, nous avons recruté 28 personnes : 24 salariés de différents profils (managers, spécialistes de ressources humaines, et autres cadres) ainsi que 4 experts académiques de la charge mentale.

L'étude a été réalisée sous forme d'entretiens semi-directifs avec chaque participant (environ une heure trente) et consistait en deux phases. Dans la première phase, les participants devaient tout d'abord donner leur propre définition de la *charge mentale*. À partir de cette définition, et pour la deuxième phase, les participants définissaient leur propre grille d'annotation avec l'aide des expérimentateurs en choisissant les étiquettes. La particularité de notre approche consiste dans le fait qu'aucun guide n'a été donné aux participants. Au contraire, ce sont eux qui devaient définir leurs propres étiquettes et les utiliser dans la tâche d'annotation. Les huit messages ont été présentés aux participants dans un ordre aléatoire. Les participants n'avaient pas accès aux annotations des autres, et ne se connaissaient pas.

La tâche d'annotation a été effectuée avec l'outil Doccano de Nakayama et al. (2018) avec l'assistance d'un expérimentateur (pour la mise en place et la configuration). En lisant un message, le participant-annotateur devait surligner (librement) les parties du texte qui révélaient pour lui une indication de la charge mentale de la personne s'étant exprimée dans le message, et attribuer à chaque segment une étiquette. L'annotation n'était pas limitée et participants pouvaient attribuer autant d'étiquettes qu'ils voulaient sur un seul segment.

La définition personnelle de la charge mentale par les participants, puis les étiquettes qu'ils ont définies, et les segments de message annotés constituent les informations principales que nous avons étudiées. D'autres informations (type de participant, durée d'annotation, etc..) seront aussi étudiées.

Résultats

Statistiques générales

Le nombre d'étiquettes utilisées varie de 5 à 15 avec une moyenne de 10 par participant. Au total, nous avons recueilli 197 étiquettes uniques pour 28 annotateurs. Ce grand nombre peut être expliqué par le fait que les annotateurs n'avaient pas de guide d'annotation et avaient la liberté de définir leurs propres étiquettes sans connaître les propositions des autres. Il s'est avéré que certaines étiquettes présentaient des similitudes, par exemple, *priorité* et *priorisation*. Toutefois, nous les avons jugées comme étant distinctes.

Parmi ces 197 étiquettes, 143 ont été proposées par un seul participant. À l'inverse, les étiquettes les plus souvent mentionnées sont *individu, temporalité* et *sens* qui ont été proposées 7, 6 et 5 fois respectivement. Cette similarité des étiquettes proposées par des utilisateurs a pu être expliquée par leur représentation proche du concept de charge mentale ainsi que par l'influence des textes à annoter.

Pour avoir une vision plus globale des dimensions considérées par les participants, nous avons utilisé un modèle d'IA générative, en l'occurrence Claude 3.5 (Anthropic, 2024), pour proposer un regroupement de toutes les étiquettes. Le prompt demande simplement un regroupement des 197 étiquettes sur la base de leur interprétation contextualisée, sans indication du nombre de clusters voulu. Le résultat est une liste de 14 thématiques présentée dans l'annexe A, avec une indication du nombre d'étiquettes correspondantes et du nombre d'annotateurs concernés. Par exemple, les étiquettes comme *échéance, urgence, temps, manque de temps* et *temporalité* ont été regroupées dans le cluster Temporalité. Dans ces 14 clusters, on trouve bien les grandes tendances de la charge mentale et notamment les dimensions qui apparaissent dans les travaux de Galy (2016) et Le Gonidec (2022) sur la charge mentale.

En revanche, les étiquettes ayant été proposées par les annotateurs eux-mêmes, il est possible que ceux-ci leur attribuent des définitions complètement différentes de l'un à l'autre, et donc employées pour décrire des éléments textuels différents.

Segments les plus annotés

Si l'on regarde maintenant les segments de textes sélectionnés, en les considérant indépendamment des étiquettes attribuées, nous avons pu observer que les participants annotent entre 13 et 248 segments au total (69 en moyenne). Les segments eux-mêmes vont d'un seul caractère (un point d'interrogation pour annoter l'emphase) à 1144 caractères (soit l'intégralité d'un message) avec une longueur moyenne de 39 caractères. Cette fluctuation était prévisible du fait de l'absence de directives et de restrictions.

Nous avons regroupé les annotations en considérant, pour chaque élément du texte (au niveau du caractère), le nombre d'annotateurs différents qui l'ont inclus dans au moins un segment annoté (donc en ignorant à la fois les étiquettes et le fait qu'un même segment peut être annoté plusieurs fois par le même participant).

Il est possible d'obtenir une visualisation des segments les plus annotés dans chaque texte, comme dans la Figure 1. *Les segments sont montrés en bleu, les plus annotés étant les plus foncés et ceux très peu annotés plus clairs*.

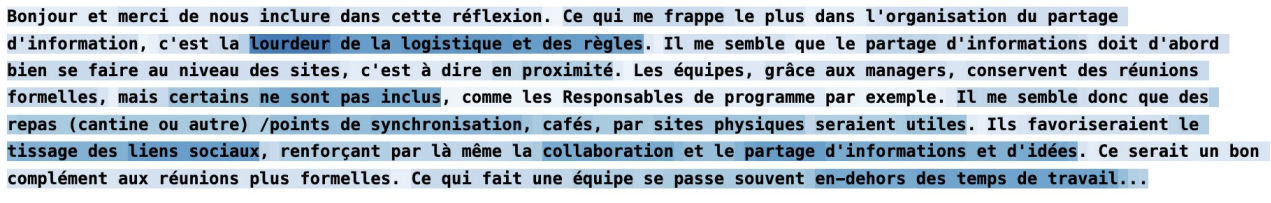


figure . 1 Exemple de texte avec marquage des zones annotées

Contexte : Collecte d'expériences sur le partage d'informations au sein de la Direction XX-XXX.

Pour une approche plus qualitative, nous avons cherché à délimiter les segments qui ont été le plus considérés par les annotateurs. En étudiant l'évolution de la taille et du nombre de segments

en fonction du seuil, et en nous basant sur le point d'inflexion, nous avons fixé un seuil à 14. Au total, sur les 8 textes, 52 segments ont été annotés par au moins 14 annotateurs; la liste complète de ces segments est citée dans l'Annexe B.

Après une première analyse de ces segments, nous pouvons remarquer que le vocabulaire contenant des mots à connotation négative attirait le plus l'attention. Les annotateurs se sont concentrés sur les éléments de texte exprimant l'inconfort, l'accablement, la surcharge et la déconnexion, des thèmes étroitement associés à une charge mentale de travail accrue. Cet aspect sera plus précisément détaillé dans la présentation.

Conclusion et futurs travaux

Cette première recherche a examiné les perceptions individuelles de la charge mentale de travail grâce à une méthode d'annotation incluant des participants de différents milieux professionnels. En donnant la possibilité aux annotateurs de créer leurs propres étiquettes et d'identifier sans contrainte des parties de texte, nous avons saisi un éventail de points de vue sur la charge mentale, et cela malgré la dispersion d'annotation et d'étiquettes (du fait que nous n'avions pas donné de consignes aux participants) et malgré la vision différente qu'ils peuvent avoir du concept de charge mentale. Comme supposé, les utilisateurs ont intégré un "socle" commun de connaissances sur le thème de la charge mentale dans le cadre de l'entreprise considérée, comme on peut le constater avec la liste des clusters en annexe A.

Nous avons voulu observer les étiquettes et les segments les plus annotés. Dans la suite du projet, nous allons aussi étudier les segments qui n'ont pas été annotés du tout par les utilisateurs. Ensuite, nous voudrions étudier plus les étiquettes et leurs relations avec les segments pour trouver les similarités maximale et minimale entre les mêmes étiquettes.

Comme nous avons enregistré toutes les passations, nous disposons de données très riches avec les définitions de la charge mentale, les reformulations, les explicitations et les commentaires lors de l'annotation. Nous allons pouvoir ainsi étudier les nuances des choix des étiquettes, la polysémie et observer les relations entre les groupes d'annotateurs et ses annotations. Nous prévoyons également de mener la deuxième phase d'expérimentation en recrutant davantage de participants avec des profils variés, non seulement au sein d'une même entreprise, mais aussi dans des autres entreprises. Pour compléter l'annotation avec un point de vue plus objectif nous allons aussi donner la même tâche d'annotation avec le même ensemble de données à des grands modèles de langue. Cela nous permettra d'intégrer les perspectives extérieures du contexte de l'entreprise et du domaine.

Dans le cadre d'un projet plus vaste, ces travaux nous servirons dans le développement d'un outil d'annotation automatique de la charge mentale de travail dans les messages de salariés. Cette nouvelle approche ouvre la voie à une meilleure compréhension des nuances linguistiques et culturelles qui influencent l'évaluation de la charge de travail mentale, tout en soulignant l'importance d'intégrer différentes perspectives pour enrichir les modèles d'analyse des données textuelles.

Bibliographie

Anthropic (2024). Claude 3.5 Sonnet.

Cabitza, F., Campagner, A., and Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI 2023)*, pages 6860–6868.

Cain, B. (2007). A review of the mental workload literature. Technical Report RTO-TR-HFM-121Part-II, Defence Research and Development Canada, Toronto.

Galy, E. (2016). Approche intégrative de la charge mentale de travail : une échelle d'évaluation basée sur le modèle ICA (Individu – Charge – Activité). In *Actes du 51e Congrès de la Société d'Ergonomie de Langue Française (SELF)*, Marseille, France.

Le Gonidec, N. (2022). *Conceptualiser et évaluer la charge mentale de salariés dans un contexte d'usage d'outils numériques : Le cas d'une entreprise de télécommunications*. PhD thesis, Université Côte d'Azur, France.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). Doccano : Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Plank, B. (2022). The 'problem' of human label variation : On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682. Association for Computational Linguistics.

Pustejovsky, J., Bunt, H., and Zaenen, A. (2017). Designing annotation schemes : From theory to model. *Handbook of Linguistic Annotation*, pages 21–72.

Valette, M. (2024). What does perspectivism mean? an ethical and methodological countercriticism. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LRECCOLING 2024*, pages 111–115.

Annexe

A. Clusters définis par Claude avec le nombre d'étiquettes et le nombre d'utilisateurs

Clusters	Annotateurs	Étiquettes
Aspects Émotionnels et Psychologiques	18	24
Relations et Interactions	18	19
Charge de Travail	17	15
Environnement et Contexte	17	17
Temporalité	17	10
Organisation et Gestion	14	16
Aspects Cognitifs et Mentaux	13	16
Contraintes et Difficultés	13	15
Équilibre et Bien-être	13	14
Capacités et Compétences	11	11
Impact et Conséquences	11	12
Processus et Actions	9	11
Adaptation et Changement	8	9
Management et Reconnaissance	5	8

B. Liste de segments annotés par au moins 14 personnes

'lourdeur de la logistique et des règles', 'ne sont pas inclus', 'points de synchronisation', 'tissage des liens sociaux', "collaboration et le partage d'informations et d'idées", 'en-dehors des temps de travail', 'on fait quoi maintenant', 'ne sais pas trop', 'peut-être dû réfléchir un peu', 'sens', 'un peu perdus', 'on ne sait pas vers où, ni pourquoi', 'malaise général', 'pas promus', 'absolument', 'garder autant', 'fallait-il pas simplifier', 'donner du sens', "dans les hiérarchies supérieures il n'y en a pas assez", 'numérique', 'les écrans ont raison de notre bien-être', 'Se reconnecter à la nature, à notre environnement, aux humains', 'moi', 'primordial', 'essentiel', 'revenir à des choses simples', 'reconnecter', 'non pas', 'des robots', 'nos émotions', 'réponse différente', 'tu as certainement un problème hormonal', 'gêne occasionnée', 'réponse différente', 'gêne', 'interrompre le travail', 'nous sommes très heureux', 'déshumanisé', 'perso', 'nauséabonde', 'très froid', 'heureuse', 'retrouver mes collègues', "perdu une part de la bonne humeur, de l'ambiance", 'échange convivial', 'informel', 'un mois avant le salon', 'et pas 3 jours avant', 'imposées', 'très importantes', 'surcharge de travail', 'laisser dans le flou'

Étude diachronique fondée sur corpus de de deux motifs phraséologiques du récit conjectural

Romain Fernandez [1,2], Iva Novakova [1] et Delphine Gleizes [2]
[1] Laboratoire LIDILEM, UGA
[2] Laboratoire Litt&Arts, UGA
romain.fernandez@univ-grenoble-alpes.fr, iva.novakova@univ-grenoble-alpes.fr,
delphine.gleizes@univ-grenoble-alpes.fr

Introduction

Cette étude a pour objectif d'analyser en diachronie deux motifs phraséologiques afin d'identifier des traits caractéristiques aux sous-genres de l'anticipation (ANT) et de la science-fiction (SF). Les deux motifs observés sont particulièrement fréquents au sein des séquences interactives et narratives de nos œuvres. Ils nous permettent donc d'analyser l'articulation de l'interaction et de la narration dans les récits conjecturaux.

ANT et SF sont deux sous-genres littéraires appartenant à la conjecture romanesque rationnelle (Versins, 1972). Souvent considérés comme synonymes (Saint-Gelais, 1999), ANT et SF font actuellement l'objet de recherches littéraires visant à repérer leurs différences socio-historiques, médiatiques et poétiques afin d'identifier des traits caractéristiques distinctifs (Bréan 2012 ; Stiénon, 2016 ; Langlet, 2006). Une première distinction s'opère à partir de l'histoire littéraire : l'anticipation naît en Europe durant la seconde moitié du XIX^{e} s. et se développe sans soutien éditorial (absence de collections). À l'inverse, la science-fiction apparait aux Etats-Unis dans les années 1920, et constitue très tôt une communauté d'amateurs. Cette SF s'exportera en France, et influencera la majorité des auteurs SF français après les années 50.

Ce travail est fondé sur une approche diachronique appliquée à la phraséologie et à la linguistique de corpus (comme récemment dans Sorba et *al*. 2024). Cette approche nous permet d'analyser les constructions des sous-genres dans leurs dynamiques (Celeyron & Sorba, 2025). Il s'agit d'un élément-clé de notre étude puisqu'il permet de les caractériser tout en se concentrant sur les différences mais aussi sur les modalités de filiation.

Corpus et méthodologie

Corpus

Nos données sont issues de deux corpus romanesques. Le premier, PhraseoAnticipation est constitué à partir de critères poétiques, éditoriaux et thématiques. Il est composé de 98 textes couvrant la seconde moitié du XIX^{e} s. et la première moitié du XX^{e} siècle. Afin d'assurer la représentativité du corpus, chaque décennie comporte entre 33 et 45% de la production écrite d'anticipation recensée par l'ANR « Anticipation »[1]. Le corpus SF fait partie du corpus littéraire

1 Pour le descriptif du projet ANR « Anticipation » (2014-2019) mené par Claire Barel-Moisan (ENS Lyon, https://anranticip.hypotheses.org/presentation)

PhraseoRom, constitué dans le cadre du projet ANR DFG PhraseoRom[2]. Les œuvres le composant proviennent des collections éditoriales labellisées SF (« Anticipation », « Le Rayon Fantastique », etc.). Totalisant 139 œuvres, le corpus s'étend de 1952 à 2014. Deux corpus de romans contemporains à nos sous-genres sont utilisés pour mesurer la spécificité d'un patron au sous-genre. Leurs caractéristiques sont consignées dans le tableau 1 Les corpus sont disponibles dans la base de données Lexicoscope V2[3].

Nom du corpus	Empan chronologique	Nombre de mots	Nombre de textes	Nombre d'auteurs
PhraseoAnticipation	1862-1939	6 758 985	98	58
PhraseoRom SF	1952-2014	10 956 882	139	34
PhraseoRom	1950-2016	107 443 867	1 123	367
CorpusContrasteXIX-XXe	1860-1939	9 230 760	88	59

table 1. : Caractéristiques des corpus utilisés

La constitution de nos corpus s'appuie sur la définition qu'en donne Rastier (2011) : il s'agit de textes intégraux, documentés, enrichis d'étiquettes, réunis d'après des considérations théoriques génériques (dans notre cas, les textes comportent un minimum de 20 000 tokens pour distinguer les nouvelles des romans) et annotés syntaxiquement via *UD* (*universal dependencies*, Abouwarda & Kraif, 2024) afin d'identifier des *motifs phraséologiques* à partir de constructions lexico-syntaxiques (CLS).

Méthodologie

Nous reprenons de la méthodologie PhraseoRom l'approche inductive (*corpus driven*), contextualiste et textométrique. Nous postulons également que la surreprésentation statistique de CLS dans les sous-genres littéraires constitue des marqueurs génériques (Siepmann, 2015). Une fois les corpus annotés et analysés syntaxiquement, une liste des collocations les plus spécifiques à chaque corpus est extraite. À cette fin, le *loglikehood ratio* (LLR, Dunning, 1993), mesure statistique permettant d'évaluer la spécificité qu'entretiennent deux cooccurrents (ici dans deux jeux de données, le corpus ANT et le corpus SF mis en comparaison), est utilisé. Les Arbres Lexico-syntaxiques Récurrents (ALR) (Kraif 2016, 2019), définis comme des associations statistiques significatives dont les unités lexicales les composant sont dépendantes syntaxiquement, sont déterminés à partir de ces collocations.

Les ALRs servent à identifier des motifs phraséologiques, définis comme étant des « cadres collocationnels » qui accueillent un ensemble d'éléments fixes et variables susceptibles d'accompagner la structuration textuelle et de caractériser des textes de genres divers (Longrée & Mellet, 2013). Ces motifs assument également des fonctions discursives (FD) au sein du discours romanesque (Adam, 2011 ; Novakova & Siepmann, 2020). L'interprétation de ces FD permet de décrire le rôle des motifs identifiés au sein de séquences romanesques ainsi que d'observer des variations/évolutions de ces derniers sur l'empan temporel couvert par nos deux corpus.

2 Projet ANR DFG PhraseoRom (2016-2020) sous la direction d'I. Novakova (UGA) et Dirk Siepmann (U. d'Osnabruck), https://lidilem.univ-grenoble-alpes.fr/node/16/axes-recherche/axe-1-descriptions-linguistiques-tal-corpus/projets-axe-1/phraseorom)

3 Le Lexicoscope a été créé par Olivier Kraif (2016) : http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/

Résultats

L'extraction des ALR les plus spécifiques au sous-genre anticipation a permis d'identifier la collocation : « répliquer_ingénieur ». Le patron du motif est défini ainsi : VerbesDicendi Det NomFonctionSocial[4]. Les variations paradigmatiques retenues pour le motif proviennent des 50 cooccurrences les plus fréquentes sur l'ensemble du corpus. Le motif ainsi décrit est hautement spécifique à ANT (LLR 419,7, il est présent dans 79 œuvres d'anticipation ; pour la science-fiction, le score est négatif : LLR -654,5, l'ALR est présent dans 44 œuvres). Le motif est également spécifique à ANT lorsqu'il est comparé à la littérature contemporaine. Le motif agrège de nombreuses extensions syntagmatiques (adjectivales : « répondit le jeune homme », adverbiales : « répliqua gravement l'ingénieur », des syntagmes prépositionnels : « d'un ton grave »). Il structure la séquence dialogale et caractérise les personnages au travers de leurs actions ou de leurs affects. Le motif assume une fonction interactive mais également infra-narrative (actions d'arrière-plan sans conséquence sur l'intrigue, Novakova & Siepmann, 2020 : 291-293). L'analyse en diachronie révèle une diminution de la fréquence à partir du XX^{e} siècle. L'on observe en revanche une nette augmentation de cette fréquence d'utilisation durant la décennie 1890-1899. Ce phénomène peut inciter à identifier un modèle de récit d'anticipation émergent sur les dernières décennies du siècle, et à caractériser ses traits distinctifs. Il peut également servir de marqueur pour interroger l'importance du récit vernien comme modèle des romans d'anticipation sur la période.

L'ALR « poser_question » est spécifique au sous-corpus science-fiction lorsque comparé à l'anticipation. (LLR 49,27, il est présent dans 132 romans science-fiction ; en revanche, son score est négatif dans anticipation : LLR -146,01, il est présent dans 72 romans). Néanmoins, le motif n'est pas un marqueur générique de la science-fiction, il est très peu spécifique au sous-corpus SF quand il est comparé à l'ensemble des sous-corpus contenus dans PhraseoRom. Les variations paradigmatiques sont plus fréquentes dans le sous-corpus ANT[5] que dans le sous-corpus SF. Il apparaît que le motif est émergent dans l'anticipation et se stabilise à partir du début du XX^{e} s. autour de la forme « poser la question ». Le motif admet plusieurs variations syntagmatiques dans SF : des adjectifs (« des questions inquiètes, personnelles, une première question, quelques questions », etc.), des syntagmes prépositionnels (« elle posa une question à peine détournée ») ou des propositions subordonnées (« la question qui lui brulait les lèvres »). Les questions dans la science-fiction concernent majoritairement l'intrigue ou les personnages. L'interrogation n'est pas une modalité d'élucidation de l'univers SF (contrairement aux personnages scientifiques de l'anticipation qui interrogent les nouveaux phénomènes) mais un ressort narratif étoffant les personnages. Le motif assume principalement une fonction discursive interactive, narrative et en de plus rares cas affective (description d'affects).

L'analyse des deux motifs révèle une évolution des séquences dialogales à partir du XX^{e} siècle. L'ALR « répliquer_ingénieur » est caractéristique du sous-genre anticipation à son origine et tend à disparaître à partir du XX^{e} s. On observe que le motif est beaucoup moins présent dans le corpus science-fiction, les traits distinctifs des personnages science-fiction dépendent moins de leurs fonctions ou de leurs genres. L'ALR « poser_question » s'impose à partir du XX^{e} siècle que le motif s'impose, à la fois en anticipation et en science-fiction, sous la forme « poser Det question ». Les questions n'attendent pas forcément de réponses dans la science-fiction, le sous-

4 Verbes Dicendi (dire|répondre|demander|répliquer|riposter) Det (l'|le|la) NomFonctionSociale (ingénieur|homme|femme|savant|capitaine|docteur, etc.)

5 Verbes (poser|soulever|adresser|formuler|avoir|permettre) Det (Cette|la|des) Question

genre délaisse l'explicite (verbalisations d'affects, explications de phénomènes) afin d'épaissir le mystère autour de son univers. L'analyse diachronique des motifs souligne ici une convergence des deux sous-genres littéraires mais laisse entrevoir deux manières de considérer la « conjecture romanesque rationnelle » (Versins). Les motifs constituent des indices permettant de dessiner les contours d'une filiation entre ANT et SF.

Références bibliographiques

Abouwarda, R. & Kraif, O. (2024). Utilisation de requêtes syntaxiques pour la terminologie : une étude de cas dans les domaines de la psychologie et de la psychiatrie. *SHS Web of conferences*, 19. https://doi.org/10.1051/shsconf/202419111005

Adam, J.-M. (2017). *Les textes : types et prototypes*. Paris. Armand Collin.

Bréan, S. (2023). Un genre encore fécond ? Devenirs de l'anticipation en France. *Belphégor*, 21-1. https://doi.org/10.4000/belphegor.5246

Bréan, S. (2012). *La Science-fiction en France. Théorie et histoire d'une littérature*. Paris. PUPS.

Celeyron, T. et Sorba, J. (2024). La phraséologie du lexique de l'armement : étude diachronique dans deux corpus romanesques outillés des 19e et 20e siècles. Travaux de linguistique, 89(2), 49-74. https://doi.org/10.3917/tl.089.0049.

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. https://doi.org/10.4000/corpus.6101.

Dunning, T. (1993). Accurate method for the statistics of surprise and coincidence. *Computational linguistics*, 19(1), 61-74.

Kraif, O. (2016). Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. *Cahiers de lexicologie*, n°108, 91.

Langlet, I. (2006). *La Science-fiction. Lecture et poétique d'un genre littéraire*. Paris. Armand Colin.

Longrée, D. & Mellet, S. (2013). Le motif : une unité phraséologie englobante ? Étendre le champ de la phraséologie de la langue au discours. *Langages*, 189, 65-79. https://doi.org/10.3917/lang.189.0065.

Novakova, I. & Siepmann, D. (Eds.). (2020). *Phraseology and Style in Subgenres of the Novel : A synthesis of Corpus and Literary Perspectives*. Palgrave Macmillan.

Rastier, F., (2011). *La mesure et le grain. Sémantique de corpus*. Paris. Champion.

Saint-Gelais, R. (1999). *L'Empire du pseudo. Modernités de la science-fiction*. Québec. Nota Bene.

Siepmann, D. (2015). A corpus-based investigation into key words and key patterns in post-war fiction. *Functions of language*. 22. 362-39 https://doi.org/10.1075/fol.22.3.03sie

Sorba, J., Kraif, O., Renwick, A., & Denoyelle, C., (2024). La constitution de corpus en diachronie longue : méthodologies, objectifs et exploitations linguistiques et stylistiques, *Corpus*, 25. https://doi.org/10.4000/corpus.8461

Stiénon, V. (2016). Une école belge de l'anticipation ?. *Textyles*, 48, 13-27. https://doi.org/10.4000/textyles.2657

Versins, P. (1972). *Encylopédie de l'utopie, des voyages extraordinaires et de la science-fiction*. Lausanne. L'Âge d'Homme.

Étude d'un chat de prévention du suicide en diachronie courte (2005-2015)

Gudrun Ledegen[1]
Nicolas Béchet[2]
Nathalie Garric[3]
Rémy Kessler[4]
Jean-Marc Leblanc[5]
Frédéric Pugnière[6]
Vanessa Thouroude[1]
[1] Université Rennes 2, Laboratoire PREFICS
[2] Université Bretagne-Sud, CNRS 6074A, Vannes
[3] Université de Nantes, Laboratoire PREFICS
[4] Université d'Avignon, LIA
[5] Université Paris Est Créteil, CEDITEC
[6] Université Bretagne-Sud, Laboratoire PREFICS

Notre analyse porte sur les 11 ans d'un corpus de chat de prévention du suicide (2005-2015, 10 millions de mots) (ANR SAPS APIPRESUI 2024-2026 (*Analyse participative et interdisciplinaire d'un chat de prévention du suicide*), porté par G. Ledegen), et plus particulièrement sur l'évolution diachronique courte des discours qui y sont tenus entre écoutants et appelants : l'étude depuis sa création en 2005, jusqu'à sa stabilisation, voire son institutionnalisation dans les années 2010, permet de révéler l'adaptation de ce dispositif, initialement conçu sous forme d'échanges téléphoniques, à un dispositif synchrone exclusivement écrit.

Dans ces interactions particulières de la communication médiée par ordinateur (CMO), où l'interaction se joue de façon « désincarnée » (Romain & Fracchiolla 2016), le face-à-face à l'écrit peut souffrir du manque de l'empathie de la voix, et l'appelant, en situation de souffrance ou de fragilité, peut se méprendre de l'intonation, la « musique » de la parole, qui accompagne le message. Par ailleurs, la production écrite n'est pas présentée au fur et à mesure de sa réalisation, mais arrive en blocs, par la sollicitation de la touche *Envoi*, sur l'écran de l'interlocuteur. L'adaptation à ce nouveau dispositif pour les écoutants, habitués à la ligne téléphonique, peut ainsi se contraster entre les années de démarrage du chat et les suivantes, où ce dispositif connaît un grand succès. Par ailleurs, le dispositif ayant été mis en place spécifiquement pour les plus jeunes, qui désertaient de fait la ligne téléphonique, des pratiques générationnelles divergentes peuvent se faire jour, entre une pratique assidue des nouvelles communications numériques tout juste émergentes en 2005, et des écoutants plus âgés pour qui l'écrit représente encore avant tout une norme graphique stable.

Notre démarche analytique prend ainsi appui sur un corpus s'inscrivant dans un genre discursif particulier, celui de la *ligne d'appel*[1] : les données présentent une cohérence de la situation d'énonciation et différents outils lexicométriques seront combinés afin de faire émerger les structures signifiantes saillantes de ce corpus, de façon quantitative et qualitative, avec un constant retour au texte (Leblanc, 2015 : 26). Le corpus a été anonymisé, tout en préservant le

1 L'étude comparative entre ses versions téléphonique, mail et chat est une perspective de travail imminente.

code d'identification de chaque écoutant, ce qui permet de réunir les discours tenus par les mêmes écoutants, et ainsi de suivre un même écoutant en interaction avec différents appelants, afin d'établir à terme une typologie des écoutants de ce corpus.

Le corpus se constituant en « séries textuelles chronologiques » (Salem, 1988 : 106), nous allons y étudier le « temps lexical » (Salem, 1988, 2021), en analysant, dans un premier temps, l'évolution des discours tenus par les écoutants de façon isolée, pour dessiner les tendances en termes de lexique utilisé, de tournures figées et de structures syntaxiques, en particulier pour ce qui est des registres adoptés (*proximité / distance*, Koch & Oesterreicher 2001). Il sera ainsi possible d'identifier quelques premières tendances dessinées sur 11 ans. Nous ferons de même pour les discours des appelants (en considérant les appelants ponctuels d'une part et les appelants récurrents d'autre part, i.e. une petite 20-aine d'appelants, appelés poly-appelants (environ 500.000 mots)), en nous axant non pas sur leur contenu variable, mais plus spécifiquement sur leurs réactions aux discours tenus par les écoutants : étonnements, interrogations sur la nature humaine ou robotique de l'interlocuteur, énervements …

Une fois ces tendances mises au jour, nous les remettrons en contexte, et aborderons le corpus dans deux directions en mobilisant une analyse quantitative et qualitative des interactions, dans le but de révéler – ou non – des évolutions diachroniques le long des 11 années du corpus :

- la première se concentrera sur différents moments et thèmes des conversations : les phases d'ouverture des conversations, où les salutations mutuelles, les invitations à se livrer, et les façons de recevoir, par les écoutants, les temps de confidence et d'argumentation du mal-être de l'appelant s'adaptent graduellement au dispositif. Par ailleurs, nous étudierons les moments où les appelants mentionnent l'inceste, le viol, ou l'acte pédophile dans un contexte émotionnel fort, avec la prise en charge par l'écoutant : diverses stratégies sont utilisées par l'appelant dont celle de contournement graphique (un v***) ou encore celle de minoration (c'était pas du viol mais de la violence physique). Du côté de l'écoutant, on observe une tendance à recentrer les propos de l'appelant sur la désignation idoine (pour déresponsabiliser la victime et pour qu'elle prenne conscience de la gravité de l'acte et des conséquences en droit…). Enfin, nous étudierons l'hypothèse d'une compétence évaluative du dispositif chez les poly-appelants en particulier, qui, à défaut de pouvoir construire de manière progressive et continue leur histoire en raison de la variable interlocuteur, s'orientent vers une forme de schématisation essentialisante de leur besoin d'aide, qui les mène à une auto-affectation stéréotypique dans une source de souffrance, laquelle fonctionne comme préconstruit (Pêcheux 1975). Les écoutants complémentairement reçoivent la parole de l'appelant par rapport à ces catégories de la souffrance, déterminants sociaux, qui imposent des appartenances collectives (vs identité individuelle).
- la deuxième abordera l'évolution des conversations non souhaitées , et importées depuis le dispositif téléphonique où ils sévissent avant tout : les « conversations troll » (Jost, 2019) et les « phonophiles » (comme ils sont désignés dans le cadre de la ligne d'écoute téléphonique, ou les « obsédés sexuels du téléphone »), sont un des principaux soucis rencontrés par les associations : ces personnes s'immiscent dans la ligne d'écoute, pour la perturber, souvent par provocation ou en inventant des problèmes. L'analyse de ces pratiques et les réactions des écoutants, et la façon dont ils tentent de révéler l'intrus, livrera des critères permettant de repérer ces conversations.

Ces analyses se construiront à l'aide de la méthodologie des « tresses chronologiques » dans *Lexico 3&5* (Salem 2021) et des AFC dynamiques dans *TextObserver* (Leblanc 2015) qui permettront de révéler et de visualiser quels aspects des discours sont particulièrement évolutifs (cf. Labbé & Monnière 2002). Les outils lexicométriques procurant les spécificités lexicales

seront travaillés en parallèle avec les spécificités structurelles révélées par la *convolution* (Vanni 2024 ; Vanni et al. 2022 ; Mayaffre & Vanni 2022) ; enfin, les *segments répétés*, les co-occurrences et les *cohortes spécifiques* (à l'aide de la *carte des sections* de Lexico5, Salem, 2021 : 41) nous guideront dans l'exploration de l'étude des figements, lesquels portent souvent la marque de la *langue de distance*.

Références

Jost, François, 2019, Troll, in Pauline Escande-Gauquié, Pauline et Naivin, Bertrand, Comprendre la culture numérique, Dunod, 170-176.

Koch, Peter & Oesterreicher, Wulfgang, 2001, « Langage oral et langage écrit », in Holtus, Günter, Metzeltin, Michel, & Schmitt, Christian (éds), Lexikon der romanistischen Linguistik. Band I, 2 : Methodologie, Tubingue, Max Niemeyer Verlag, 584-627.

Labbé, D. et Denis, M., 2002, Essai de stylistique quantitative. Duplessis, Bourassa et Lévesque », dans Annie Morin et Pascale Sébillot, VIe Journées Internationales d'analyse des données textuelles, Saint-Malo, 13-15 mars 2002, Rennes, IRISA-INRIA, no 2, 561-569.

Lebart, L. & Salem, A., 1994, Statistique textuelle, Paris, Dunod.

Lebart, L., Pincemin, B. & Poudat, C., 2019, Analyse des données textuelles, Québec, Presses de l'Université du Québec.

Leblanc, J.-M., 2015, Proposition de protocole pour l'analyse des données textuelles : pour une démarche expérimentale en lexicométrie, Nouvelles perspectives en sciences sociales, 11 (1), 25-63.

Ledegen, G. & Wagener, A., 2020, « Nous ne doutons pas de votre souffrance » : analyse pragmatique et sociolinguistique du nous de distanciation dans un chat de prévention du suicide, Corpus, 21, Garric, N., Ledegen, G. et Pugnière-Saavedra, F. (Dirs), 'Dispositifs numériques et dévoilement de soi', https://journals.openedition.org/corpus/4977.

Ledegen, G., 2019, 'Bonsoir. Je vais mal'. La difficile expression du dévoilement de soi et de l'empathie dans un chat de prévention du suicide, in Abécassis, M., Block, M., Ledegen, G., Peñalver Vicea, M., Le grain de la voix dans le monde anglophone et francophone, Oxford, Peter Lang, 193-214.

Longhi, J. & Salem, A., 2018, Approche textométrique des variations du sens, Proceedings of the 14th International Conference on Statistical Analysis of Textual Data, Rome, Universitalia, 452-458.

Mayaffre, D. & Vanni, L., 2024, Usages linguistiques des éléments supplémentaires dans l'Analyse factorielle des correspondances. JADT, Anne Dister; Dominique Longrée, Jun 2024, Bruxelles, Belgique. pp.613-623. hal-04647313

Romain, Christina & Fracchiolla, Béatrice, 2016, Violence verbale et communication numérique écrite : la communication désincarnée en question. Les cahiers de praxématique, Emotions en contexte numérique, 66, https://ff10.4000/praxematique.4263ff

Salem, André, 1988, Approches du temps lexical, Statistique textuelle et séries chronologiques, Mots. Les langages du politique, 17, 105-143.

Salem, André, 2021, Le temps lexical, *Histoire & mesure* [En ligne], XXXVI-2 | 2021, http://journals.openedition.org/histoiremesure/14804.

Vanni, L., 2024, Hyperbase Web. (Hyper)Bases, Corpus, Langage. *Corpus*, 2024, 25, ⟨10.4000/corpus.8770⟩. ⟨hal-04523479⟩

Vanni, L., Guaresi, M. & Magri, V., 2022, Convolution et marqueurs multidimensionnels. Description des représentations genrées dans un corpus de films français. 16th International Conference on Statistical Analysis of Textual Data (JADTS 2022), Jul 2022, Naples, Italie.

Étudier l'Italien populaire à l'aide de la linguistique de corpus

Bianco Francesco[1]
[1] Université Grenoble Alpes
francesco.bianco@upol.cz

Introduction: la recherche sur l'italien populaire

Le panorama linguistique de l'Italie est d'une richesse exceptionnelle, offrant une grande variété de formes linguistiques rarement observée dans un espace géographique aussi restreint. D'un point de vue sociolinguistique, le répertoire linguistique italien présente une complexité remarquable, synthétisée par Gaetano Berruto (2020) dans sa célèbre « architecture » de l'italien contemporain, qui demeure un modèle de référence.

Parmi les variétés qui composent ce paysage, les linguistes ont mené de nombreuses recherches approfondies sur ce que l'on appelle l'italien populaire ou italien des semi-cultivés (italiano popolare, italiano dei semicolti ; dorénavant : IP), une variété diastratique substandard de l'italien, définie diversement par les chercheurs : par exemple, comme une variété d'italien utilisée par des locuteurs peu instruits (De Mauro 1970), ou encore comme une variété imparfaitement acquise par des individus dont la langue maternelle est un dialecte italien (Cortelazzo 1972). L'IP fut un objet de débat intense dans les études linguistiques italiennes entre le début des années 1970 et la première moitié des années 1990.

La plupart des chercheurs (Bruni 1978, 1984 ; Cortelazzo 1972 ; Bartoli Langeli 2000 ; Marazzini 2004, 2006) considèrent que l'IP n'existe qu'à l'écrit. Ils n'envisagent pas ou sous-estiment son existence en tant que variété orale. L'IP serait donc un italien écrit comportant des éléments de langue parlée, introduits par le faible niveau d'instruction de l'auteur. Ce scénario a récemment été remis en question par certains chercheurs (Berruto 2014, 2020 ; Fresu 2016 ; Guerini 2016 ; Lubello 2018) qui suggèrent l'existence d'un IP parlé. Toutefois, en raison du manque de données, aucune analyse détaillée ni projet de recherche conséquent n'a encore été réalisé.

En ce qui concerne la dimension historique, tous les chercheurs ne s'accordent pas sur la présence actuelle de l'IP (Cortelazzo 2001 ; Lepschy 2002 ; D'Achille 2010). La vision traditionnelle associe cette variété à une époque où la population italienne était majoritairement peu instruite et dialectophone (dont la langue maternelle était uniquement un dialecte). Pour ces auteurs, l'IP serait typique de la période allant de la seconde moitié du XIXe siècle à la première moitié du XXe siècle ; au cours des 50 à 70 dernières années, d'autres variétés, comme le néo-standard italien — c'est-à-dire une évolution de l'italien standard traditionnel —, auraient pris sa place (Sobrero 2005).

Malgré de nombreuses initiatives de recherche, l'organisation de colloques dédiés et la publication d'un grand nombre d'articles et d'ouvrages, certaines lacunes importantes persistent, à mon avis, pour deux raisons principales :

- Dans le cadre des corpus de la langue italienne, aucun outil n'a été spécifiquement conçu pour explorer cette variété (Chiari 2012 ; Ježek & Sprugnoli 2023) ;
- Dans les travaux sur ce sujet, et plus généralement dans les recherches sur les variétés de

l'italien, les chercheurs ont principalement appliqué des critères d'évaluation qualitatifs, parfois même « impressionnistes » (Berruto 2020) ; les méthodes quantitatives, les humanités numériques (HN) et les outils de data science sont encore largement négligés, bien que quelques travaux récents commencent à les explorer (Sprugnoli et al. 2019).

Une archive

Le but de ce travail est de présenter les résultats d'APID (building an Archive of Popular Italian Documents), initiative de recherche soutenu par le projet GATES (Grenoble ATtractiveness and ExcellenceS), financé par l'Agence Nationale de la Recherche (ANR) au titre du programme France 2030 avec la référence ANR-22-EXES-0001. APID vise à combler partiellement les lacunes évoquées ci-dessus, en concevant et réalisant le tout premier corpus dédié à l'IP. Ce corpus servira également de base de données pour des expérimentations, qui prolongeront les travaux entamés dans le cadre de projets antérieurs (Bianco & Ghezzi 2005b). L'atteinte de ces deux objectifs constituera une première étape dans la réduction du vide scientifique autour de l'IP, et ouvrira de nouvelles perspectives pour la recherche sur la variation linguistique (italienne).

Dans le cadre de ce projet, seule la variété écrite de l'IP est prise en compte. Les documents inclus dans le corpus appartiennent principalement aux « formes primaires de l'écrit » (Folena 1985) : lettres, cartes postales, journaux intimes, notes, etc. Certains ont déjà été collectés et partiellement analysés dans mes recherches précédentes, sans pour autant avoir été publiés systématiquement (Bianco 2013, 2016). D'autres documents sont issus d'ouvrages publiés (par exemple Di Stasio 1991) ou d'archives (comme l'Archivio Diaristico Nazionale à Pieve Santo Stefano, Italie). Les documents sont encodés légèrement selon la norme XML-TEI, afin de permettre leur annotation et leur traitement (recherche automatique, visualisation, etc.).

Un des défis majeurs concerne la présence de données sensibles et privées (noms de personnes, lieux, entreprises, montants financiers, etc.) dans ces textes. Ces données sont pseudonymisées, afin de respecter à la fois la réglementation en vigueur et les droits des personnes concernées, tout en maintenant une bonne lisibilité des documents. Ce type d'anonymisation a déjà été appliqué avec succès par d'autres groupes de recherche (Clemenzi et al. 2023), et est intégré dès les premières étapes de la chaîne de traitement. Aucune donnée brute (scans ou transcriptions non anonymisées) ne sera diffusée.

Concernant l'encodage XML-TEI, il est léger, principalement centré sur les métadonnées de base (source, date, genre, etc.), les caractéristiques textuelles principales (sauts de ligne et de page, titres, etc.) ainsi que les entités nommées, en vue de leur pseudonymisation.

La sortie prévue du projet prendra la forme d'un ensemble de documents anonymisés encodés en XML-TEI, mis à disposition selon deux modalités : (a) via une interface web de recherche simplifiée ; (b) via une API dédiée à des usages plus avancés. Cette approche orientée HN vise à garantir plusieurs avantages :

- Offrir aux chercheurs une large gamme de possibilités d'analyse, notamment l'intégration facilitée de données issues de sources diverses (ce qui est souvent difficile ou impossible aujourd'hui) ;
- Assurer une pérennité maximale à la production scientifique du projet ;
- Maintenir les coûts et délais du projet dans les limites prévues.

De plus, le corpus est conçu comme un outil flexible et minimaliste, apte à être enrichi, réutilisé ou intégré dans d'autres projets. Le code source sera publié dans un dépôt accessible au public.

Recherches passées et futures

Jusqu'à présent, l'IP a été étudié par des chercheurs isolés, principalement à travers des méthodes traditionnelles et qualitatives. Il semble qu'aucun projet collectif ni corpus-based/driven n'ait été mené précisément sur cette thématique (Chiari 2012). Le projet APID s'inscrit dans un programme de recherche visant à combler cette lacune (Bianco & Ghezzi 2025a).

Dans un travail précédent (Bianco & Ghezzi 2025b), mon équipe et moi avons présenté les résultats de nos premières expérimentations sur l'IP, en appliquant des principes, outils et technologies relevant de l'intelligence artificielle (IA). À ce stade initial, notre objectif était de vérifier si un classifieur fondé sur l'apprentissage automatique (machine learning, ML) pouvait identifier des textes non standard et de poser les jalons pour une future application plus large de l'IA à la variation linguistique italienne.

La constitution de ce nouveau corpus permettra, dans les prochaines annés, d'approfondir cette recherche.

Références bibliographiques

Bartoli Langeli, A. (2000). La scrittura dell'italiano. Bologna: il Mulino.

Berruto, G. (2010). Varietà. EncIt, p. 1550-1553, <https://www.treccani.it/enciclopedia/varieta_%28Enciclopedia-dell%27Italiano%29/>.

Berruto, G. (2014). Esiste ancora l'italiano popolare? Una rivisitazione. Dall'architettura della lingua italiana all'architettura linguistica dell'Italia. Saggi in omaggio a Heidi Siller-Runggaldier, éd. par P. Danler & C. Konecny. Frankfurt am Main: Peter Lang, p. 277-290.

Berruto, G. (2020). Sociolinguistica dell'italiano contemporaneo. Roma: Carocci.

Bianco, F. (2013). Le lettere dei migranti irpini fra italiano, dialetto e lingua straniera. Variante et varieté – Variante e varietà – Variante y variedad – Variante und Varietät. Actes du Vie Dies Romanicus Turicensis, Zurich, 24-25 juin 2011, éd. par C. Albizu et al.. Pisa : ETS, p. 101-117.

Bianco, F. (2016). Dalla periferia al centro: un secolo di storie di irpini emigrati in Nord America (1911-2010). Études romanes de Brno, 37, 2, p. 133-143, <https://digilib.phil.muni.cz/handle/11222.digilib/135895>.

Bianco, F. & Ghezzi, S. E. (2025a). Tracce di italiano popolare nel parlato contemporaneo. Italiano LinguaDue, 16, 2, p. 652–672, <https://riviste.unimi.it/index.php/promoitals/article/view/27874>.

Bianco, F. & Ghezzi, S. E. (2005b). Beyond ChatGPT: AI applications to Italian language varieties. Artificial Creativity. Looking at the Future of Digital Culture, éd. par A. Micalizzi. Cham: Springer, p. 7-20.

Bruni, F. (1978), Traduzione, tradizione e diffusione della cultura: contributo alla lingua dei semicolti. Quaderni storici, 13, p. 523-554.

Bruni, F. (1984). L'italiano. Elementi di storia della lingua e della cultura. Torino: UTET.

Cerruti, M. & Ballarè, S. (2021). ParlaTO: corpus del parlato di Torino. Bollettino dell'Atlante Linguistico Italiano (BALI), 44 [2020], p. 171-196.

Chiari, I. (2012). Corpora e risorse linguistiche per l'italiano. Stato dell'arte, problemi e prospettive. Italienisch, 68, p. 90-105.

Clemenzi, L. et al. (2023). Testi in maschera: nuovi strumenti per la sicurezza e l'analisi linguistica di corpora giuridici. Umanistica Digitale, 16, p. 1-32.

Cortelazzo, M. (1972), Avviamento critico allo studio della dialettologia italiana. Lineamenti di italiano popolare. Pisa: Pacini.

Cortelazzo, M. A. (2001). L'italiano e le sue varietà: una situazione in movimento. Lingua e stile, 36, 3, p. 417-430.

D'Achille, P. (2010). Italiano popolare. EncIt, p. 723-726, <https://www.treccani.it/enciclopedia/italiano-popolare_(Enciclopedia-dell%27Italiano)/>.

De Mauro, T. (1970). Per lo studio dell'italiano popolare unitario. Lettere da una tarantata, éd. par A. Rossi. Bari: De Donato, p. 43-75.

Di Stasio, G. (éd.)(1991). Ti sono scritto questa lettera. Le lettere che gli emigranti non scriveranno più. Milano: Mursia.

EncIt = Enciclopedia dell'italiano, éd. par R. Simone. Roma: Istituto della enciclopedia italiana, 2010.

Folena, G. (1985). Premessa. Le forme del diario [Quaderni di retorica e poetica, 2], éd. par G. Folena, p. 5-10.

Fresu, R. (2016). L'italiano dei semicolti. Manuale di linguistica italiana, éd. par S. Lubello. Berlin-Boston: De Gruyter, p. 328-350.

Guerini, F. (2016). Il corpus ParVa. Rilevanza per la ricerca e applicazioni didattiche. Italiano e dialetto bresciano in racconti di partigiani, éd. par F. Guerini. Roma: Aracne, p. 9-38.

Ježek, E. & Sprugnoli, R. (2023). Linguistica computazionale. Introduzione all'analisi automatica dei testi. Bologna: il Mulino.

Lepschy, G. (2002). Popular Italian: Fact or Fiction? Mother Tongues and Others Reflections on the Italian Language. Toronto: University of Toronto Press, p. 49-69.

Lubello, S. (2018). Scritture da lontano: semicolti campani e sondaggi dal corpus MeTrOpolis. Testi e linguaggi, 12, p. 229-237.

Marazzini, C. (2004). Breve storia della lingua italiana. Bologna: il Mulino.

Marazzini, C. (2006). La storia della lingua italiana attraverso i testi. Bologna: il Mulino.

Sobrero, A. (2005). Come parlavamo e come parliamo. Spunti per una microdiacronia delle varietà dell'italiano. Gli italiani e la lingua, éd. par F. Lo Piparo & G. Ruffino. Palermo: Sellerio, p. 209-220.

Sprugnoli, R. et al. (2019). Analysing the Evolution of Students' Writing Skills and the Impact of Neo-standard Italian with the help of Computational Linguistics. Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-It 2018, éd. par E. Cabrio et al. Torino: Accademia University Press.

Évaluer automatiquement la complexité syntaxique des productions orales enfantines à l'aide d'un outil en ligne

Loïc Liégeois [1], Christine da Silva-Genest [2], Christophe Benzitoun[3], Caroline Masson[4] et Christophe Parisse[5]
[1] LRL, Université Clermont Auvergne ; [2] DevAH, Université de Lorraine ; [3] ATILF, Université de Lorraine ; [4] CLESTHIA, Université Sorbonne Nouvelle ; [5] MODYCO, Université Paris Nanterre
loic.liegeois@uca.fr, christine.da-silva-genest@univ-lorraine.fr ; christophe.benzitoun@univ-lorraine.fr ; caroline.masson@sorbonne-nouvelle.fr ; cparisse@parisnanterre.

L'évaluation du langage enfantin : état des lieux

L'évaluation de la complexité du langage enfantin est une thématique récurrente dans le domaine de l'acquisition du langage. Puisque l'une des caractéristiques principales du langage de l'enfant est qu'il se complexifie au fur et à mesure de son développement, l'évaluation de cette complexité est souvent vue comme primordiale dans le but, par exemple, de situer un enfant particulier par rapport à une cohorte. Le seul âge de l'enfant n'est pas suffisant pour déterminer son stade de développement linguistique. Ainsi, la longueur moyenne des énoncés (MLU, Brown, 1973) a très vite été utilisée comme un indice de complexité des énoncés enfantins et reste encore aujourd'hui l'étalon de mesure du développement du langage le plus utilisé. Toutefois, on sait que cette mesure, n'est plus fiable au-delà de l'âge de 48 mois (Klee & Fizgerald, 1985 ; Klee et al., 1989 ; Blake et al., 1993). Pour autant, le langage enfantin continue de se complexifier au-delà de cet âge : la proportion d'énoncés complexes s'accroit largement (Diessel, 2004). De nombreux instruments de mesure du développement du langage ont donc été proposés comme la *clausal density* (Berman, 1996) ou le *Complex Syntax Type-Token Ratio* (Witkowska et al., 2022). Il existe également des mesures plus globales du développement morphosyntaxique telles que l'*Index of Productive Syntax* (Scarborough, 1990) ou le *Developmental Sentence Scoring* (Lee, 1974) pour l'anglais. Malgré l'utilisation de ces mesures pour évaluer le langage d'enfants plus âgés, il est difficile de s'accorder sur une mesure syntaxique pertinente et corrélée à un degré de complexité du langage.

De plus, la méthodologie de recueil des données à l'origine du calcul des indices de complexité est un facteur important à prendre en compte. En effet, elle influence directement la complexité syntaxique des énoncés observés. Par exemple, les tâches de récit et de restitution d'histoire induisent, chez des enfants de 5 à 7 ans, un langage plus complexe qu'une conversation classique recueillie en situation naturelle (Westerveld & Vidler, 2016 ; Westerveld et al., 2004).

En dépit de son intérêt, les mesures de complexité syntaxique du langage enfantin spontané sont peu utilisées et laissent la place à l'utilisation de tests de langage qui privilégient des situations de production très contraintes et dirigées. Notre objectif est double. Il s'agit tout d'abord, dans la continuité de nos précédents travaux, de présenter la mesure que nous avons développée pour évaluer la complexité des productions spontanées d'enfants âgés de plus de 4 ans. Celle-ci fera l'objet d'une validation empirique. Ensuite, notre communication s'articulera autour de la présentation de l'outil en ligne que nous développons et qui permettra le calcul automatique de la mesure de complexité à partir du dépôt d'une transcription par l'utilisateur.

Évaluer la complexité du langage enfantin au moyen de données naturelles et spontanées

Dans le cadre d'interventions orthophoniques, les professionnelles portent un intérêt certain à l'évaluation d'un langage enfantin spontané. En effet, les situations naturelles d'interaction engendrent une meilleure implication de l'enfant et permettent à l'orthophoniste de mieux saisir ses compétences. Cela induit une intervention orthophonique plus efficace (Rodi, 2017). Comparées aux indicateurs fournis par les résultats à des tests de langage standardisés, les mesures issues de l'analyse du langage spontané offrent une forte fiabilité, une validité plus robuste et une meilleure sensibilité (Guo & Eisenberg, 2015; Pavelko & Owens, 2017).

Cependant, les orthophonistes se trouvent confrontées à plusieurs types de difficultés lorsqu'il s'agit d'évaluer le degré de complexité d'énoncés enfantins produits en situation d'interaction. Tout d'abord, il n'existe pas de mesure claire, standardisée et étalonnée de la complexité. Si les spécialistes ont une bonne intuition de ce qui est complexe chez un enfant, il n'existe pas à ce jour d'indices fiables et partagés afin de la mesurer scientifiquement. Par ailleurs, les comparaisons entre sujets au développement atypique (ou entre enfants au développement typique et atypique) s'avèrent difficiles car les mesures ne sont pas standardisées (da Silva et al., 2023). De plus, la mise en œuvre d'une évaluation fondée sur des interactions spontanées avec l'enfant s'avère particulièrement chronophage pour les orthophonistes (Klatte et al., 2022) et peu adaptée à leur pratique professionnelle.

C'est pour répondre à ces problématiques que nous avons souhaité développer, à la suite de nos précédents travaux, un outil permettant une mesure de la complexité syntaxique à partir de l'analyse automatique de corpus de productions spontanées de jeunes enfants. Cet outil se veut être facile d'utilisation, accessible en ligne, reposant sur des indices fiables et permettant une comparaison des données, tout en tenant compte des aspects éthiques et règlementaires.

Définition d'un indice de complexité

Pour définir un indice de complexité syntaxique, deux pistes ont été explorées. Une première piste « M16 (règles ad-hoc) » s'appuie sur les caractéristiques de la grammaire de l'oral (Blanche-Benveniste,1990) et des productions enfantines orales spontanées. Nous avons créé une typologie du degré d'élaboration langagière en retenant 16 critères de complexité (présence d'un sujet nominal, d'adjectifs épithètes ou de temps verbaux spécifiques). Ensuite, cette typologie est appliquée automatiquement aux corpus de productions enfantines par des méthodes de traitement automatique du langage. Dès qu'un des 16 critères retenus apparait dans l'énoncé d'un enfant, celui-ci est considéré comme complexe. Chaque critère apparaît donc comme un marqueur révélateur de la présence d'une caractéristique linguistique exceptionnelle, rare, ou classiquement considérée comme associée à un discours complexe par des spécialistes de l'acquisition du langage.

Une deuxième piste « Apprentissage TTK » est basée sur l'utilisation de mesures lexico-syntaxiques automatiques réalisées au moyen d'un outil développé dans le projet ANR TextToKids (Blandin et al., 2020). Ces mesures sont appliquées au corpus Colaje (Morgenstern et Parisse, 2012) ainsi qu'au corpus EVALANG (da Silva-Genest, Christmann et Willaime, 2023 ; da Silva-Genest, Masson et Guigourès, 2023). Ce dernier corpus est constitué de productions spontanées d'enfants, avec et sans trouble du développement du langage (TDL), âgés de 5 à 7 ans (Liégeois et al., 2024, da Silva-Genest et al., 2023). Parmi l'ensemble des mesures, nous retenons les 20 les plus efficaces, c'est-à-dire celles qui sont corrélées avec l'âge de l'enfant. Ces 20 mesures sont ensuite utilisées pour apprécier le nombre d'indices

linguistiques inhabituellement élevés dans un énoncé (plus que le 80ème quantile). À la différence de la méthode M16, la méthode TTK n'induit donc pas de biais tirés de l'expertise des spécialistes et se fonde uniquement sur une méthode automatique d'apprentissage non-supervisé. Avec cette méthode, un énoncé est considéré comme complexe si au moins un des indices apparaît inhabituellement élevé dans l'énoncé, comparé aux données du corpus d'apprentissage.

Application des deux méthodes à un corpus de productions enfantines

Le premier de nos objectifs est de présenter les résultats obtenus suite à l'application des méthodes précédemment décrites à un corpus de productions naturelles d'enfants âgés de 5 à 7 ans. Pour le moment, les méthodes M16 et TTK ont été appliquées aux productions de 6 enfants, avec et sans TDL, recueillies dans des situations de jeu et de récit. Ces enfants sont issus du corpus EVALANG, et les données issues de leurs productions n'ont pas participé à l'apprentissage TTK. L'ensemble des données regroupent environ 1500 énoncés. L'apprentissage TTK a été réalisé avec deux jeux de données : le reste du corpus EVALANG et le corpus COLAJE. Les résultats obtenus ont été comparés avec un codage manuel réalisé par les auteurs de l'étude. En d'autres termes, nous avons comparé la classification binaire des énoncés (simple / complexe) effectuée manuellement, automatiquement via la méthode M16 et automatiquement via la méthode TTK. Les résultats sont présentés dans la table ci-dessous.

Mesure (méthode de traitement)	Rappel	Précision	F1
M16 (règles ad-hoc)	0.70	0.58	0.63
Apprentissage TTK (corpus XXX)	0.57	0.52	0.54
Apprentissage TTK (corpus COLAJE)	0.60	0.49	0.54
Combinaison de méthodes (M16 et TTK)	0.85	0.50	0.63

La mesure de Rappel indique quelle proportion d'énoncés complexes est bien repérée. La valeur 0.85 indique ainsi que 85% des énoncés complexes ont été identifiés. La mesure de Précision informe sur le nombre d'énoncés correctement identifiés comme complexes. La valeur 0.50 indique que 50% des énoncés annotés comme complexes le sont effectivement. La meilleure mesure de qualité des algorithmes, F1, tient quant à elle compte du rappel et de la précision.

Nous pouvons observer que la méthode *ad hoc* M16 est la meilleure sur notre corpus de test. Il est surtout intéressant que les deux méthodes produisent des résultats qui ne se recouvrent pas exactement. En les combinant, nous récupérons davantage d'énoncés complexes (85%) au prix d'une méthode moins précise qui attrape plus d'énoncés non complexes. En effet, 50% des énoncés identifiés comme complexes ne sont pas considérés comme tels lors de l'annotation manuelle effectuée par des experts.

Présentation de l'outil

Notre second objectif est de présenter l'outil en ligne dcpx (Parisse, 2025) mis à disposition de la communauté (https://dcpx.ortolang.fr/fr/). Cet outil permet d'obtenir un indice de complexité des productions d'un enfant à partir de la transcription d'un enregistrement (10-15min). Cette transcription pourra être directement chargée par l'utilisateur ou bien généré par un outil de transcription automatique. Notre cahier des charges était le suivant : cet outil devait être simple d'utilisation, accessible en ligne, transparent au niveau des calculs effectués et respectueux des principes éthiques et législatifs. Dans cette partie de la communication, nous nous focaliserons

donc à la fois sur les problématiques techniques et éthiques que posent la mise à disposition d'un tel outil, avant d'effectuer une courte démonstration de son utilisation.

Références bibliographiques

Berman, R. A. (1996). "Form and function in developing narrative abilities: the case of 'and'," in Social Interaction, Context, and Language: Essays in Honor of Susan Ervin-Tripp, eds D. Slobin, J. Gerhardt, A. Kyratzis, and J. Guo (Mahwah: Lawrence Erlbaum), 343–367.

Blake, J., Quartaro, G., & Onorati, S. (1993). Evaluating quantitative measures of grammatical complexity in spontaneous speech samples. Journal of Child Language, 20(1), 139–152. https://doi.org/10.1017/S0305000900009168

Blanche-Benveniste, C. (1990). Le français parlé. Éditions du CNRS.

Blandin A., Lecorvé, G., Battistelli, D. et Étienne, A. (2020). Age Recommendation for Texts. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 1431–1439, Marseille, France. European Language Resources Association.

Brown, R. W. (1973). A first language: The early stages. Harvard University Press.

Diessel, H. (2004). The Acquisition of Complex Sentences. Cambridge University Press.

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.

Klatte I., van Heugten V., Zwitserlood R., et Gerrits E., 2022, « Language sample analysis in clinical practice: speech-language pathologists' barriers, facilitators, and needs », Language, Speech, and Hearing Services in Schools, 53(1), p.1-16.

Klee, T., Schaffer, M., May, S., Membrino, I., & Mougey, K. (1989). A comparison of the age-MLU relation in normal and specifically language-impaired preschool children. Journal of Speech and Hearing Disorders, 54, 226–233.

Klee, T., & Fitzgerald, M. D. (1985). The relation between grammatical development and mean length of utterance in morphemes. Journal of Child Language, 12, 251–269.

Lee, L. L. (1974). Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians. Northwestern University Press.

Liégeois, L., da Silva, C., Benzitoun, C., Masson, C. et Parisse, C. (2024). Méthodologie(s) de segmentation des transcriptions de l'oral en vue de la création d'une mesure d'évaluation des productions verbales enfantines. Colloque International La linguistique de l'oral spontané à travers les langues : création, annotation et analyse de corpus, segmentation du discours. INALCO, 23-24 mai 2024, Paris.

Morgenstern, A., & Parisse, C. (2012). The Paris Corpus. Journal of French Language Studies, 22(Special Issue 01), 7–12. https://doi.org/10.1017/S095926951100055X

Parisse, C. (2025). dcpx. [Logiciel]. https://dcpx.ortolang.fr/fr/

Pavelko S. L. et Owens J. R. E., 2017, « sampling utterances and grammati- cal analysis revised (SUGAR): New normative values for language sample analysis measures », Language, Speech, and Hearing Services in Schools, 48, p. 197-215.

Rodi M., 2017, « Les interactions du quotidien comme ressources pour l'évalua- tion des capacités langagières », GLOSSA, 120, p. 75-98.

Scarborough, H. S. (1990). Index of productive syntax. Applied Psycholinguistics, 11, 1-22.

da Silva-Genest, C., Christmann, A.-L. et Willaime, P., 2023. Ressources en acquisition et pathologie de l'acquisition du langage. 11ème Journée de Linguistique de Corpus, Juillet 2023, Grenoble, France.

da Silva, C., Liégeois, L., Masson, C., Benzitoun, C. et Le Mené, M., 2023. Création d'un référentiel lexical à partir des productions verbales d'enfants à développement typique et atypique en situation de jeu. 11e Journée de Linguistique de Corpus, Université de Grenoble, 3-7 Juillet 2023, Grenoble. [hal-04153646]

da Silva-Genest, C., Masson, C. et Le Mené Guigourès, M., 2023. Évaluer les compétences morphosyntaxiques et discursives d'enfants d'âge scolaire : analyse d'une situation de jeu libre. Études de linguistique appliquée : revue de didactologie des langues-cultures et de lexiculturologie, 2023, 210 (2), pp.179-195. [10.3917/ela.210.0053].

Westerveld, M. F., Gillon, G. T., & Miller, J. F. (2004). Spoken language samples of New Zealand children in conversation and narration. Advances in Speech Language Pathology, 6(4), 195–208. https://doi.org/10.1080/14417040400010140

Westerveld, M. F., & Vidler, K. (2016). Spoken language samples of Australian children in conversation, narration and exposition. International journal of speech-language pathology, 18(3), 288–298. https://doi.org/10.3109/17549507.2016.1159332

Witkowska, D., Lucas, L., Jelen, M. B., Kin, H., & Norbury, C. (2022). Development of complex syntax in the narratives of children with English as an Additional Language and their monolingual peers. PsyArXiv. https://doi.org/10.31234/osf.io/96dx5

Exploration de l'organisation du lexique mental des adultes sans troubles à travers un corpus d'associations verbales

ELIE-DESCHAMPS Juliette [1], TERRIER Emeline [2]
[1] Laboratoire CeReS, Université de Limoges (87000)
[2] Orthophoniste, Saint-Michel (16470)
juliette.elie-deschamps@unilim.fr, emelineorthoterrier@gmail.com

Introduction

Un adulte sans trouble possède environ 50 000 à 70 000 mots de sa langue maternelle (Bogliotti, 2012; Ferrand, 1994; Van der Linden, 2006). Ces chiffres pourraient faire penser que choisir les bons mots quand nous souhaitons parler représente une tâche importante et nécessite de solliciter des mécanismes complexes. Or, un locuteur lambda commet très peu d'erreurs de production, environ une pour mille mots produits (Bogliotti, 2012; M. Rossi & Peter-Defare, 1998), ce qui laisse entendre que le stockage des mots en mémoire à long terme est bien organisé.

Collins et Quillian (1969) se sont penchés sur cette organisation et ont proposé de représenter le lexique mental en un vaste réseau de connexions. Chaque unité lexicale stockée en mémoire est reliée à d'autres par des liens sémantiques plus ou moins forts, entrainant des degrés d'activation différents entre ces unités.

De nombreuses études ont porté sur la structuration du lexique chez des enfants sans trouble (Bassano, 1998). Des normes d'associations verbales ont également pu être établies dans cette population et chez l'adulte sans trouble à partir de tâches d'association libre simple (Ferrand & Alario,1998 ; Ferrand, 2001 ; De La Haye, 2003 ; Tarrago et al., 2005 ; Duscherer & Mounoud, 2006 ; Bonin et al., 2013) ou de tâches de fluence sémantique (Dubois, 1983 ; Marchal & Nicolas, 2003 ; Karousou et al., 2023; Dorchies et al., 2024).

Malgré cette variété de recherche, il est toujours difficile de se représenter l'organisation des unités de la langue que possède un individu au sein de son lexique mental. Nous proposons ici de vous présenter les résultats d'une étude portant sur l'exploration du lexique mental à travers une tâche d'association libre continuée, utilisée dans une précédente étude (Elie-Deschamps, 2019) et ayant pu montrer des perspectives intéressantes.

Méthodologie et corpus

Sujets

Cinquante sujets ont participé à l'étude (25 hommes et 25 femmes) dont l'âge moyen est de 24;4 ans (étendue de 18;2 ans à 30;7 ans). Les niveaux d'études et les profils professionnels varient en fonction des sujets (la majorité sont des étudiants ou des actifs).

Stimuli

Le matériel comprend 20 noms communs concrets appartenant à 10 catégories sémantiques différentes (table 1). Au sein de chaque catégorie sémantique, nous avons sélectionné un

élément prototypique (désignés par un * dans le tableau), c'est-à-dire l'un des plus représentatifs de sa catégorie (Niklas-Salminen, 2015), et un élément peu fréquent (Dubois, 1983; Marchal & Nicolas, 2003).

Catégorie sémantique	Mot-inducteur	Catégorie sémantique	Mot-inducteur
Animal	CHIEN* TAUPE	Véhicule	VOITURE* SCOOTER
Fruit	POMME* MIRABELLE	Meuble	TABLE* PENDERIE
Légume	CAROTTE* ASPERGE	Instrument	GUITARE* CORNEMUSE
Métier	BOULANGER* ÉBOUEUR	Outil	MARTEAU* TRONÇONNEUSE
Sport	FOOTBALL* SURF	Vêtement	PANTALON* CALEÇON

table 1. : Mots-inducteurs sélectionnés selon leur catégorie sémantique et leur fréquence

La tâche proposée aux sujets était une tâche orale d'association libre continuée. Celle-ci consiste à présenter oralement un mot-inducteur au sujet en lui demandant d'associer spontanément tous les mots qui lui viennent à l'esprit. L'association est dite libre puisqu'il n'y a aucune contrainte de production, et continuée puisque le sujet n'est pas limité dans son nombre de réponses (Moliner & Lo Monaco, 2017). Elle se distingue ainsi d'autres études basées sur une tâche écrite d'association libre simple (Bonin et al., 2013; De La Haye, 2003; Ferrand, 2001).

Corpus

Un total de 4850 réponses a été fourni par nos participants et constitue donc notre corpus pour cette présente étude. Chacune de ces réponses a fait l'objet d'une analyse qualitative et quantitative.

Résultats

Les résultats de notre étude permettent de mettre en évidence une différence significative entre le nombre total de réponses apportées pour les mots prototypiques et pour les mots peu fréquents (IC [9 ; 69], $p = 0.015$). Les sujets ont donné au total 38 réponses de plus pour les prototypes (261 contre 223). Une différence significative est également observée concernant l'étendue lexicale (IC [3 ; 30], $p = 0.023$), avec en moyenne 17 types de plus pour les prototypes (89 contre 72). L'étendue lexicale correspond au nombre de réponses différentes au sein du corpus (un mot donné 15 fois compte pour 1 seul type).

Pour ce qui concerne le type d'association, notre corpus montre que, quel que soit le type d'item (prototypique ou peu fréquent), les réponses associatives en lien avec la fonctionnalité (36%), la description (23%) et la taxonomie (20%) sont les plus nombreuses. Nous pouvons constater, grâce à la figure 1, que la relation d'hyponymie n'est pas du tout représentée dans le corpus des mots peu fréquents. La proportion d'associations de type partie-tout et fonction est deux fois plus importante pour les prototypes. A l'inverse, nous retrouvons davantage de réponses en lien avec un élément associé et le contexte spatial pour les mots peu fréquents.

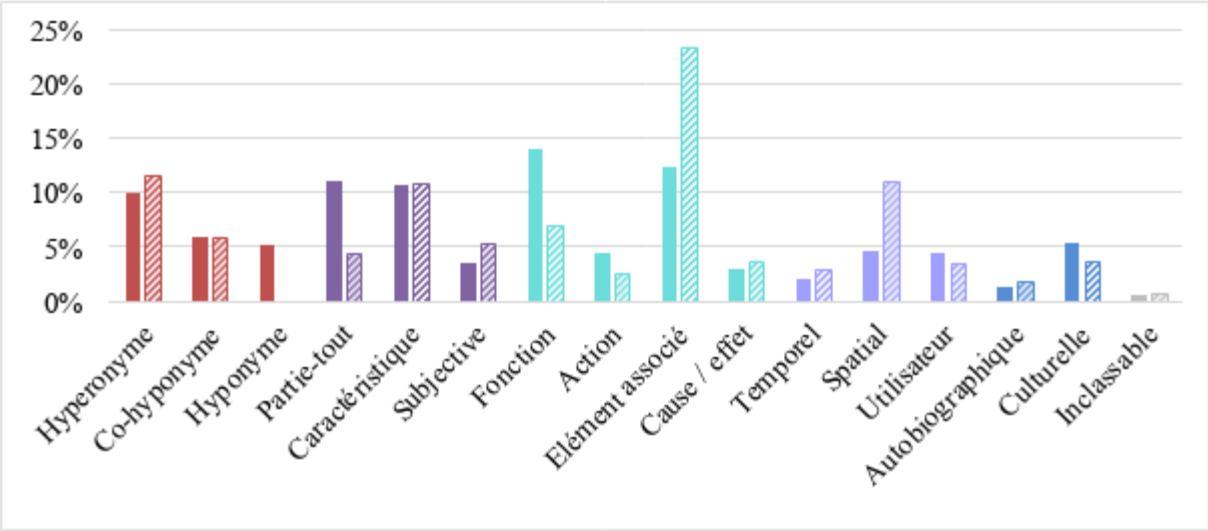


figure . 1 Comparaison des types d'association en fonction de la fréquence des mots-inducteurs

Les associés de 1er, 2e et 3e ordre représentent les trois associés ayant les pourcentages de force associative les plus élevés. La force associative est mesurée par le nombre d'occurrences d'une même réponse au sein du corpus d'un mot-inducteur. Nous avons pu constater que ces trois associés-là correspondent à des mots renvoyant en majorité à la relation d'hyperonymes (23%) et à des éléments associés (21%).

L'analyse de la réponse donnée en premier par les sujets, appelée 1er associé, met en évidence l'intérêt d'une tâche d'association libre continuée, puisque, dans notre étude, le 1er associé le plus fréquemment donné n'est identique à l'associé de 1er ordre que pour 13 mots-inducteurs sur 20 (65%).

Conclusion

Nos résultats permettent de montrer que les types d'association varient en fonction du mot inducteur, de son type (prototypique ou peu fréquent) mais aussi en fonction du participant. Cependant, la tâche d'association libre continuée utilisée ici met en évidence que les réponses venant spontanément en premier à l'esprit des sujets (1ers associés) sont également celles qui sont les plus fréquemment données par l'ensemble des sujets (associés de 1er, 2e et 3e ordre). Le corpus recueilli montre également que le lexique est fortement organisé autour de la taxonomie, en particulier l'hyperonyme.

Références bibliographiques

Bassano, D. (1998). L'élaboration du lexique précoce chez l'enfant français: structure et variabilité [Structure and variability in French early lexical development]. *Enfance, 4,* 123–153. https://doi.org/10.3406/enfan.1998.3120

Bogliotti, C. (2012). Les troubles de la dénomination. *Langue française*, *174*(2), 95-110. https://doi.org/10.3917/lf.174.0095

Bonin, P., Méot, A., Ferrand, L., & Bugaïska, A. (2013). Normes d'associations verbales pour 520 mots concrets et étude de leurs relations avec d'autres variables psycholinguistiques. *L'Année psychologique*, *113*(1), 63-92. https://doi.org/10.3917/anpsy.131.0063

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240-247. https://doi.org/10.1016/S0022-5371(69)80069-1

De La Haye, F. (2003). Normes d'associations verbales chez des enfants de 9, 10 et 11 ans et des adultes. *L'Année psychologique*, *103*(1), 109-130. https://doi.org/10.3406/psy.2003.29627

Dorchies, F., Muchembled, C., Adamkiewicz, C., Godefroy, O., & Roussel, M. (2024). Investigating the cognitive architecture of verbal fluency : Evidence from an interference design on 487 controls. *Frontiers in Psychology*, *15*, 1441023. https://doi.org/10.3389/fpsyg.2024.1441023

Dubois, D. (1983). Analyse de 22 catégories sémantiques du français : Organisation catégorielle, lexique et représentation. *L'Année psychologique*, *83*(2), 465-489. https://doi.org/10.3406/psy.1983.28477

Duscherer, K., & Mounoud, P. (2006). Normes d'associations verbales pour 151 verbes d'action. *L'Année psychologique*, *106*(3), 397-413.

Elie-Deschamps, J. (2019). « Évaluer la structuration lexicale via une tâche de fluence sémantique étendue dans les Troubles du Spectre de l'Autisme », dans S. Topouzkhanian et G. Hilaire-Debove (dir.). *Troubles du Spectre de l'Autisme : recherche et orthophonie*. Ortho Edition.

Ferrand, L. (1994). Accès au lexique et production de la parole : Un survol. *L'Année psychologique*, *94*(2), 295-311. https://doi.org/10.3406/psy.1994.28759

Ferrand, L. (2001). Normes d'associations verbales pour 260 mots « abstraits ». *L'Année psychologique*, 101(4), 683-721. https://doi.org/10.3406/psy.2001.29575

Ferrand, L., & Alario, F.-X. (1998). Normes d'associations verbales pour 366 noms d'objets concrets. *L'Année psychologique*, 98(4), 659-709. https://doi.org/10.3406/psy.1998.28564

Karousou, A., Economacou, D., & Makris, N. (2023). Clustering and Switching in Semantic Verbal Fluency : Their Development and Relationship with Word Productivity in Typically Developing Greek-Speaking Children and Adolescents. *Journal of Intelligence*, 11(11), 209. https://doi.org/10.3390/jintelligence11110209

Marchal, A., & Nicolas, S. (2003). Normes de production catégorielle pour 38 catégories sémantiques : Étude sur des sujets jeunes et âgés. *L'Année psychologique*, *103*(2), 313-366. https://doi.org/10.3406/psy.2003.29639

Rossi, M., & Peter-Defare, E. (1998). *Les lapsus ou comment notre fourche a langué*. PUF.

Tarrago, R., Martin, S., De La Haye, F., & Brouillet, D. (2005). Normes d'associations verbales chez des sujets âgés. *European Review of Applied Psychology*, *55*, 245-253. https://doi.org/10.1016/j.erap.2005.05.001

Van der Linden, E. (2006). Lexique mental et apprentissage des mots. *Revue française de linguistique appliquée*, *XI*(1), 33-44. https://doi.org/10.3917/rfla.111.44

Exploration des phrases préfabriquées en contexte polémique dans les discussions en ligne

Céline Poudat[1], Agnès Tutin[2], Mathilde Dargnat[3] et Maëlle Debard[1]
[1] Laboratoire Base Corpus Langages, Université Nice Côte d'Azur
[2] Laboratoire LIDILEM, Université Grenoble Alpes
[3] Laboratoire ATILF, Université de Lorraine
celine.poudat@univ-cotedazur.fr, agnes.tutin@univ-grenoble-alpes.fr, mathilde.dargnat@univ-lorraine.fr, maelle.debard@univ-cotedazur.fr

Introduction

La présente communication se propose d'explorer les phrases préfabriquées polémiques dans les interactions en ligne, et en particulier dans les échanges entre Wikipédiens. La discussion Wikipédia est un genre intrinsèquement argumentatif, dans lequel les Wikipédiens échangent sur l'article en cours, justifient et argumentent leurs contributions. Si le désaccord est accepté, et même considéré comme inévitable (Jemelniak, 2014), le conflit est quant à lui peu toléré. La frontière entre désaccord et conflit étant souvent difficile à déterminer clairement, nous nous étions concentrés dans des travaux précédents sur une caractérisation lexicale et pragmatique des marqueurs du désaccord et de l'attaque verbale (Poudat & Chandelier, 2024). Observer les phrases préfabriquées dans ces interactions (désormais PPIs) nous intéresse à plus d'un titre, tant d'un point de vue descriptif que du point de vue de la dynamique des discours (Chandelier et al, 2024). Dans quelle mesure les interactions conflictuelles dans Wikipédia sont-elles préfabriquées ? Contiennent-ils plus de PPIs polémiques, entendues comme discours disqualifiant un discours cible (Vlad, 2017 : 60) – ce qui les distingue notamment des attaques *ad hominem ?*

Afin d'apporter des éléments de réponse à ces questions, nous avons mené une première étude exploratoire des PPIs polémiques dans les discussions Wikipédia. En partant d'une liste de candidates PPIs à haut potentiel polémique dont nous avons évalué la présence dans le corpus WikiDiscussions (Section 1), nous avons extrait un échantillon d'occurrences pour chaque PPI polémique, analysé avec un modèle d'annotation des PPIs (2.). Nous présentons enfin quelques éléments de résultats, que nous développerons davantage dans notre présentation orale (3.).

Méthodologie de sélection des PPIs polémiques dans Wikipédia

Sélection des candidats PPIs polémiques

Notre intérêt porte principalement sur les PPIs qui peuvent avoir une fonction polémique, dans la mesure où leur visée pragmatique dominante est de « discréditer l'adversaire, et le discours qu'il est censé tenir » : « tous les énoncés polémiques […] se *focalisent* sur le discours adverse, et sa dénégation » (Kerbrat Orecchioni, 1980 : 11-12). Cette fonction polémique se manifeste fréquemment par l'emploi de structures syntaxiques et de phrases préfabriquées comme *tu plaisantes ! no comment. Non mais, je rêve ! si tu le dis* (Vlad, 2017). Bien que la tonalité générale du corpus Wikidiscussion soit en général assez policée et courtoise (les phrases

préfabriquées de salutation et de remerciement y sont très productives), ce corpus n'est pas exempt d'échanges plus vifs qui relèvent du discours polémique, où les locuteurs, au-delà du désaccord argumenté, disqualifient les propositions d'autres contributeurs.

A partir des listes constituées au LIDILEM dans le cadre du projet PREFAB[1], un ensemble de PPIs polémiques a été retenu en tenant compte de leur fréquence dans Wikipédia. La présente communication se fonde sur les 31 candidates PPIs polémiques, listés dans la table 1.

Corpus

Le corpus analysé est un sous-ensemble du *EFG WikiCorpus* (Ho-Dac, 2024), qui rassemble l'ensemble des discussions Wikipédia autour des articles, produites jusqu'à 2019. Il comporte 60,2 millions de tokens, analysé sur le plan syntaxique et décrit suivant un sous-ensemble de métadonnées: durée des échanges, nombre d'intervenants, nombre de posts, schéma conversationnel, titre de la page …

Le corpus est en ligne sur le Lexicoscope (Kraif, 2019), ce qui permet d'effectuer des requêtes ciblées, exploitant la syntaxe et les métadonnées. Il est également possible d'effectuer des recherches sur les cooccurrents fréquents. Les 31 PPIs polémiques retenues ont été extraites et une annotation fine a été réalisée sur un échantillon aléatoire d'au moins 50 PPIs.

Grille d'annotation des PPIs

Une grille d'annotation des PPIs, utilisée dans le cadre du projet PREFAB, a été proposée afin de décrire les propriétés syntaxiques, sémantiques et interactionnelles de ces expressions (Table 3 en annexe). Cela permet de mieux comprendre le fonctionnement des expressions dans l'interaction, en observant comment elles interviennent dans l'échange des tours de parole.

Cette grille est appliquée aux exemples extraits du corpus, comme dans (1). Dans cet exemple, on observe que la PPI *tu plaisantes ?* est déclenchée par la reprise du discours d'un autre contributeur ici reproduit (déclenchement par hétérorépétition), segment sur lequel porte la PPI (autoportée antérieure).

1. « Par tes remarques bien souvent portées uniquement sur la personne de Sartorius » :**tu plaisantes** bien sûr ? Avant d'avancer des choses pareilles lis attentivement l'intégralité de cette pdd. « j'ai bien précisé que je voulais que (…) » [wikidisc_prefabv3_D_01_.xml_691,(2007) {title}]

Résultats et perspectives

L'annotation des 31 candidats PPIs polémiques laisse entrevoir des résultats intéressants, tant du point de vue de la stabilité de la forme que de ses usages en contexte. La table 1 permet ainsi d'observer l'ambiguïté formelle des PPIs polémiques retenues. La forme la moins polémique est *c'est ça*, très largement utilisée pour valider plutôt que pour disqualifier un discours (*oui, c'est ça !* 90% des formes analysées) tandis que pour certaines formes, l'extraction des formes comme PPIs doit être retravaillée (par exemple, différencier *tu parles !* et *les exemples dont tu parles !* reste souvent complexe).

La plupart des PPIs relevées sont utilisées au sein (54%) ou au début (27,25%) d'un tour de parole. Les PPIs polémiques sont rarement utilisées en finale, ce qui contrevient parfois à

1 https://prefab.hypotheses.org/ Un grand merci à Daciana Vlad, qui nous a aidées à défricher cette liste.

l'intuition. Ainsi *point barre* n'est jamais utilisé en finale d'un tour de parole, peut-être pour éviter le trop catégorique. Seuls *no comment* et *j'ai déjà donné* peuvent clôturer un tour de parole de manière polémique.

100% PPIs polémiques	80-99%	<60%
et alors, point barre, c'est une blague, ben voyons, no comment, si le coeur vous en dit, grand bien te fasse, tiens donc	c'est n'importe quoi, laisse tomber, on s'en fout, c'est pas la question, je rêve, cherchez l'erreur, c'est à voir, si tu le dis, ça commence à bien faire, c'est l'hôpital qui se fout de la charité, tu rigoles, c'est du vent, alors quoi, j'hallucine, et puis quoi encore, on aura tout vu, c'est la meilleure, elle est bien bonne, j'ai déjà donné	c'est ça, comme par hasard, ça suffit, tu parles

table 1. : table 1 : PPIs polémiques retenus

Le relevé systématique des cooccurrents d'une forme permet en outre de mettre en évidence des régularités combinatoires, comme pour *je rêve*, précédé par un *non mais* dans 42,86% des occurrences, et suivi par un *ou quoi* dans 21,43% des cas.

Lemme	mais (a)	mais (a) là (p)	non mais (a)	non mais (a) ; là (p)	ou quoi (p)	Total général
je rêve	21,43%	7,14%	42,86%	7,14%	21,43%	100,00%
Total général	**21,43%**	**7,14%**	**42,86%**	**7,14%**	**21,43%**	**100,00%**

table 2. : table 2. : combinatoire de *je rêve* avec cooccurrents antérieurs (a) et postérieurs (b)

Enfin, le croisement des analyses avec les métadonnées des Wikis laisse entrevoir des régularités intéressantes. Ainsi, les PPIs polémiques sont rares dans les fils monologaux (moins de 15% des cas), et plus de 50% des occurrences relevées adviennent dans les multilogues (> 3 contributeurs).

Perspectives

Nous espérons avoir présenté l'intérêt d'une analyse des PPIs polémiques, même si notre étude n'en est encore qu'à son stade exploratoire et connaîtra différents ajustements, notamment pour apprécier les variations possibles d'une PPI. La mise en évidence, par l'étude de ces expressions, d'une dimension polémique, peut servir l'analyse de la structure polyphonique et argumentative du discours, qu'il se présente comme monologue ou polylogue.

Nous envisageons notamment d'observer les séquences et les cooccurrences de PPIs polémiques dans un fil de discussion, tout en étudiant la manière dont les PPIs se combinent (par exemple *ben voyons + c'est une blague*), et en systématisant l'étude des mots du discours accompagnateurs, en particulier les marqueurs associés à la concession, relation de discours argumentative polyphonique (par ex. *mais, quand même, seulement, pourtant, néanmoins*, etc.), les marqueurs associés à l'accord, qui permettent valider avant de l' « attaquer » une proposition antécédente (par ex. *oui, d'accord, certes, tout à fait*, etc.), ou encore les marqueurs associés à la surprise, émotion propice à l'introduction d'un désaccord face à l'événement inattendu l'ayant déclenchée (par ex. *ah, ben, tiens*, etc.).

Annexe

Lemme et acception	**Lemme**	Tu plaisantes
	Acception	Tu plaisantes (incrédulité)
Syntaxe	**Type phrase**	clausative
	Modalité	interrogative
	Ponctuation	?
	Propriétés syntaxiques	In situ (type d'interrogative)
	Expansion	
	Modifieurs	
Combinatoire	**Cooccurrents**	bien sûr (postposé)
Propriétés interactionnelles	**Position**	Médiane
	Déclenchement	Déclenché par hétérorépétition
	Portée	Autoportée (ante)
Fonctions	**Fonction globale**	Expressive et réactive
	Fonction spécifique	Exprimer son incrédulité Exprimer son désaccord

table 3. : table 3. : Grille d'analyse des occurrences des PPI

Références bibliographiques

Chandelier, M., Tutin, A., Etienne, C., & Poudat, C. (2024). Expression formulaire du désaccord dans les réunions de travail et les fils de discussion Wikipédia9e Congrès Mondial de Linguistique Française, Jul 2024, Lausanne (CH), Suisse.

Ho-Dac, L.-M. (2024). Building a comparable corpus of online discussions on Wikipedia, *in* Poudat, C, Lüngen, H., Herzberg, L. (eds). Investigating Wikipedia: Linguistic corpus building, exploration and analysis, Studies in Corpus Linguistics, 121, John Benjamins Publishing Company, pp.12-44.

Kerbrat-Orecchioni, C. (1980). *Le discours polémique*. Presses universitaires de Lyon.

Kraif, O. (2019). Explorer la combinatoire lexico-syntaxique des mots et expressions avec le LEXICOSCOPE. *Langue française*, *203*(3), 67-82.

Poudat, C. et Chandelier, M. (2024). « Disagreements and conflicts in Wikipedia talk pages » in Poudat, C., Lüngen, H. and Herzberg, L. (eds.), Investigating Wikipedia: Linguistic corpus building, exploration and analysis. Studies in Corpus Linguistics, 121, John Benjamins Publishing Company, pp.205-234.

Vlad, D. (2017), *Pour une « grammaire » du polémique. Étude des marqueurs d'un régime discursif agonal*, Cluj-Napoca / Oradea, Presa Universitară Clujeană & Editura Universităţii din Oradea.

Exploring the use of English metalanguage in early modern sources: Focus on orthography and variation

Vahid Asadidehziri[1], Angela Andreani[2] , Daniel Russo[3]
[1] University of Milan, Department of Languages, Literatures, Cultures and Mediations
[2] University of Milan, Department of Languages, Literatures, Cultures and Mediations
[3] University of Insubria, Department of human sciences and territorial innovation

Introduction

In the evolution of human language, one key characteristic is the ability to talk about language itself. This ability, often described as metalinguistic competence, linked with metalinguistic awareness, enables speakers not only to communicate but also to reflect on and evaluate the structures and features of language, such as spelling, grammar, vocabulary, and usage (Baddeley & Voeste, 2012). Metalanguage evidently plays a significant role in this process, allowing individuals to characterise and negotiate the rules of language both formally and informally. Even though the use of metalanguage is deeply embedded in everyday and scholarly communication, its historical trajectory, especially how it developed in early stages of the English language, remains underexplored. During the Early Modern English period, 1500 to 1700, English underwent significant transformations, not only in its grammar and orthography but also in the intellectual frameworks used to describe language and linguistic phenomena. Discussions about language itself proliferate in texts ranging from grammar treatises and spelling books to philosophical writings. In this context, tracing spelling variation in key metalinguistic terms can shed light on how linguistic awareness and reflective practices developed at a time before orthographic norms had stabilised, thereby highlighting broader shifts in early linguistic analysis. The Renaissance revival of classical learning, followed by Enlightenment ideas about knowledge and systematisation, created new opportunities for the emergence of new ways of thinking and talking about language.

However, English premodern linguistics and terminology have been treated in a piecemeal manner in the fields of ESP/LSP, general linguistics and the history of ideas. The marginal place of premodern English is evident in seminal collections on the metalanguage of linguistics (Colombat et Savelli 2001; Orioles, Bombi, Brazzo 2012), and in the International Handbook of Special-Language and Terminology Research, where only one of three essays on premodern English addresses linguistics. While Walmsley revisits grammatical terminology historically (2022), a broader understanding of the development of English linguistic metalanguage remains lacking. Building on recent work by Andreani and Russo (2023), this study aims to fill this gap by analysing the development of metalinguistic terminology and concepts in a small corpus of historical texts (see description below). This terminology will be analysed through the lens of orthographic change in Early Modern English texts. In fact, despite the historical and linguistic importance of this period, there has been limited investigation into how English spelling evolved from 1500 to 1800 (Condorelli, 2023). Detailed insights into spelling changes have often remained underexplored due to the high cost of manual analysis. This research therefore seeks to implement approaches to automatically examine whether and to what extent spelling

variation occurred during the period in question, taking into account differences across authors and historical phases.

Corpus and Methodology

Corpus

The MetaLing corpus was developed using Alston's multivolume Bibliography of the English Language from the Invention of Printing to the Year 1800 as a foundational reference (Alston, 1965-2011). The project focused on a manageable selection of texts. The initial emphasis was on language variation, based on Volume 9 of Alston's Bibliography, covering English and Scottish dialects as well as cant and vulgar English. In addition, further sources were drawn from Volume 3, covering miscellaneous works on the English language and Volume 6, covering works on pronunciation. The final corpus was organized into three subcorpora, each with a balanced number of texts. Glossaries and untranslated manuscripts were excluded in this first phase due to time constraints and limited transcription availability. Some short previously untranscribed texts were processed using the AI-powered handwriting and text recognition programme Transkribus (https://www.transkribus.org/). Corpus creation involved building a database of source metadata (e.g., author, date, format), Downloading XML files from the Oxford Text Archive, Extracting relevant text segments and uploading the corpus to Sketch Engine with metadata for linguistic analysis.

Keyword and Term Extraction

To explore patterns of metalinguistic discourse, we conducted a preliminary keyword and term extraction process. Twenty-one language-related keywords were manually and randomly selected from a small subset of the corpus. These keywords represent core linguistic concepts and terminology relevant to the study of Early Modern English. Examples include syllable, speak, spelling, linguistic, grammar, and language, and common verbs such as have and has, among others. Using Natural Language Processing (NLP) techniques, we automatically identified and retrieved variant forms of these keywords across the corpus. These variants were not predefined; instead, they were identified through tokenisation and pattern-matching algorithms using regular expressions, allowing for the automated detection of all attested spelling variants across the corpus. The aim was to trace how such terms were used and evolved over time. This process also served as a foundational step toward building a more comprehensive inventory of metalanguage in historical English texts. In combining computational and manual approaches to linguistic research, this study addresses the potential of bridging historical linguistics with digital practices for the analysis of historical texts (McGillivray, 2020).

Results

The analysis of spelling variation across Early Modern English texts reveals notable patterns in the usage and evolution of language-related terminology. The data shows that words such as "have," "has," "speech," "word," and "language" appear in a variety of spellings across different authors and time periods, reflecting the lack of standardized orthography during the 16th and 18th centuries. For example, the verb "have" frequently occurs in the form "haue" in earlier texts, such as those by Dekker and Vaughan, while "hath" and "hast" appear as common variants of "has." Similarly, "speech" is often written as "speache". The word "word" appears in multiple forms, including "worde" and "wordes," highlighting morphological variation in addition to spelling differences. Overall, the preliminary findings point to dynamic and evolving orthographic practices, with clear evidence of linguistic variation across time and

authorship. This underscores both the richness and the instability of spelling conventions in Early Modern English.

In relation to metalinguistic expression, terms like “language,” “discourse,” “tongue,” and “letter” are widely distributed across the corpus and appear consistently in texts from various decades. These recurring terms point to sustained metalinguistic interest among authors of the period, as may be expected, considering the characteristics of the corpus (see corpus description above). In later texts, more specialized terms such as “topick,” “syllable,” and “semantick” begin to appear, suggesting a gradual enrichment of linguistic vocabulary and concepts.

References

Alston, R. C. (1965-2011). A bibliography of the English language from the invention of printing to the year 1800. Scolar press.

Andreani, A., & Russo, D. (2023). Mapping the history of language-related terminology in English (1500-1700): A corpus-based collocate approach. Linguistica e filologia 43, pp. 151-174.

Baddeley, S., & Voeste, A. (2012). Orthographies in early modern Europe. De Gruyter Mouton.

Colombat, B. et M. Savelli. (2001). Métalangage et terminologie linguistique: actes du colloque international de Grenoble. Peeters.

Condorelli, Marco (Ed.). (2023). Advances in Historical Orthography, c. 1500–1800. Cambridge University Press.

McGillivray, B. (2020). Computational Methods for Semantic Analysis of Historical Texts. In K. Schuster and S. Dunn (Eds.), The Routledge Handbook of Research Methods in Digital Humanities (pp. 305-324) Routledge.

Orioles, V., R. Bombi e M. Brazzo. (2012). Proceedings of the first workshop on the metalanguage of linguistics. Il Calamo.

Walmsley, J. (2022). Language turned back on itself. Growth and structure of the English metalanguage. In S. Coffey (Ed.), The History of Grammar in Foreign Language Teaching (pp. 173-190), Amsterdam University Press.

Figure de l'ennemi dans les discours politico-médiatiques russophones contemporains et leur réception sur les réseaux socionumériques : étude sur corpus

Polina Ukhova
PRAXILING (UMR5267), Université Paul-Valéry Montpellier3
polina.ukhova@univ-montp3.fr

Cette communication interroge la dynamique des représentations de l'ennemi et du héros dans les discours russophones contemporains, à partir de l'analyse de la construction idéologique du *russkiy mir* (« monde russe ») et de sa diffusion dans les sphères politico-médiatique et socionumérique. Depuis la montée en puissance du discours nationaliste au sein du régime de Vladimir Poutine, la Russie se positionne au centre d'un récit civilisationnel revendiquant une singularité identitaire, culturelle et spirituelle, opposée à l' « Occident collectif» fréquemment décrit comme manipulé et décadent — une représentation qui sert notamment à légitimer l'invasion de l'Ukraine, présentée comme une nation sous emprise américaine et menaçante pour la sécurité nationale. Cette prétendue décadence est associée, entre autres, à l'adoption par les sociétés occidentales d'un agenda libéral, incarné notamment par la reconnaissance des droits des personnes LGBT+, l'accès à l'avortement ou encore la légalisation du mariage pour tous. Ce discours s'insère également dans le cadre d'une vision géopolitique élargie : Poutine, en tant que leader des BRICS, promeut un *monde multipolaire*, où chaque pays doit préserver ses traditions, sa culture et son identité. Cette vision est partagée par plusieurs pays membres du groupe, qui contestent l'hégémonie occidentale et appellent à une réorganisation de l'ordre mondial. Ce narratif prend racine dans une longue histoire, notamment intellectuelle et littéraire. Les écrivains russes – de Tolstoï à Soljenitsyne – ont contribué à forger une vision du monde fondée sur des valeurs morales, la spiritualité chrétienne orthodoxe, le courage du peuple et la mission salvatrice de la Russie. Leurs œuvres chantent depuis des siècles la grandeur, la force et la vocation particulière de la Russie comme « Terre Sainte », nourrissant une mémoire collective et un imaginaire national qui structurent encore aujourd'hui les discours publics.

Cette continuité discursive lie étroitement passé et présent dans la réactivation d'un passé héroïque, notamment dans la lutte contre le fascisme, dans l'apparition de l'oukase (décret) *« Sur les valeurs traditionnelles »* correspondant à un impressionnant ensemble de préceptes: le patriotisme, le civisme et le service de la patrie, le travail comme pratique constructive, l'adoption d'idéaux moraux élevés, la solidité de la famille et la priorité du spirituel sur le matériel, l'esprit du collectif, l'entraide et le respect mutuel, la mémoire historique, la continuité générationnelle et, enfin, l'unité des peuples de la Russie, ceci étant souvent opposé aux « mœurs occidentales » perçues comme décadentes. Dans cette perspective, la Russie est présentée comme une civilisation à part, un « pays-civilisation » porteur de valeurs éternelles et non négociables.

Le concept de *valeurs traditionnelles*, devenu une formule interdiscursive, circule aujourd'hui dans les discours officiels médiatiques, mais aussi dans les discours ordinaires sur les réseaux socionumériques. Pour les russophones soutenant la politique de l'État, il fonctionne comme un véritable idéologème structurant. Il contribue à renforcer la polarisation entre « nous » (le monde russe) et « eux » (l'Occident), en ressuscitant une opposition ancienne, dans laquelle les États-Unis demeurent l'ennemi numéro un. L'ennemi, dans ce contexte, est discursivement construit à travers un procédé de dramatisation (Charaudeau, 2006), comme un « mal absolu », une figure maléfique. À cette figure s'oppose celle du héros, défini comme le garant des valeurs traditionnelles, prêt à affronter l'ennemi dans un combat quasi mythique. Cette dichotomie permet une mobilisation émotionnelle et idéologique, tant dans les discours que dans leurs réceptions.

Nous proposerons une étude qualitative comparative de ces deux figures (ennemi et héros), à partir de deux corpus russophones et francophones collectés entre février 2022 et mars 2025, afin d'analyser la manière dont elles sont construites, mises en récit et reçues par les locuteurs. Cette analyse mettra en lumière la circulation des imaginaires politiques dans les discours ordinaires et leur rôle dans la promotion d'un monde multipolaire.

Le corpus A est principalement constitué de discours politico-médiatiques. Parmi la multitude d'énoncés diffusés sur divers supports médiatiques, nous avons sélectionné ceux qui font l'objet d'une publicisation marquée dans l'espace public, dans la mesure où ils permettent de constater que des notions telles que *russkiy mir* (« monde russe »), *monde multipolaire* ou *valeurs traditionnelles* fonctionnent comme de véritables référents sociaux. Il s'agit d'énoncés produits — à l'oral comme à l'écrit — par des acteurs politiques, puis repris et relayés par les médias : articles de presse, publications sur les canaux officiels sur TELEGRAM, extraits d'interventions télévisées diffusées sur YouTube, etc. Il s'agit de plus de 500 textes de différente longueur (allant de 250 mots à 250000 mots).

À titre illustratif, nous citerons quelques exemples tirés du Corpus A (extraits du discours prononcé par président Poutine lors des forums de Valdaï (en 2023)[1] et celui qu'il a adressé à l'Assemblée fédérale (29/02/2024[2]) dont les scripts sont disponibles sur le site officiel du Kremlin). Dans ces discours, le héros est présenté comme porteur d'une mission morale et civilisationnelle :

> (1) « мы – оплот традиционных ценностей » [...] (Trad. Nous sommes le bastion des valeurs traditionnelles) ;
> (2) « мы выбираем жизнь » [...] (Trad. Nous choisissons la vie) ;
> (3) « поддержка семей с детьми – наш фундаментальный нравственный выбор » [...] (Trad. Le soutien aux familles avec enfants est notre choix moral fondamental) ;
> (4) « для нас важны ценности любви, доверия в купе с культурой, историей и нравственными устоями » [...] (Trad. Pour nous, les valeurs d'amour, de confiance, associées à la culture, l'histoire et les fondements moraux, sont essentielles) ;
> (5) « мы со всей энергией и доброй волей включились в процессы строительства нового справедливого мироустройства » [...] (Trad. Nous participons avec toute notre énergie et notre bonne volonté à la construction d'un nouvel ordre mondial plus juste) ;

1 Disponible sur Youtube (URL : https://www.youtube.com/watch?v=Uqyv_Go4WBo).

2 Disponible sur Youtube (URL : https://www.youtube.com/watch?v=7ntAdLacD7o).

(6) « мы ведем праведную борьбу за суверенитет и безопасность » [...] (Trad. Nous menons un combat juste pour la souveraineté et la sécurité) ;
(7) « нам помогает наша вековая сплоченность, наша всепобеждающая сила » [...] (Trad. Ce qui nous aide, c'est notre unité séculaire, notre force irrésistible) ;

L'ennemi, quant à lui, est représenté comme une entité malveillante, hypocrite et destructrice :

(8) « коллективный Запад с колониальными запашками » [...] (Trad. L'Occident collectif aux relents coloniaux) ;
(9) « с привычкой разжигать конфликты, сдерживать развитие РФ » [...] (Trad. Habitué à attiser les conflits et à freiner le développement de la Russie) ;
(10) « это люди, которые не прошли тяжёлые испытания и забыли, что такое война » [...] (Trad. Ce sont des gens qui n'ont pas traversé d'épreuves difficiles et ont oublié ce qu'est la guerre) ;
(11) « русофобы, придерживаются идеологии расизма, национального превосходства и исключительности » [...] (Trad. Ce sont des russophobes qui adhèrent à une idéologie de racisme, de supériorité nationale et d'exceptionnalisme) ;
(12) « они ответственны за разрушение норм морали, института семьи » [...] (Trad. Ils sont responsables de la destruction des normes morales et de l'institution de la famille) ;
(13) « толкают целые народы к вымиранию и вырождению, а мы выбираем жизнь » [...] (Trad. Ils poussent des peuples entiers à l'extinction et à la dégénérescence, tandis que nous choisissons la vie) ;
(14) « США взяли курс на гегемонию » [...] (Trad. Les États-Unis ont pris le chemin de l'hégémonie) ;
(15) « прибегают к запугиванию этническими чистками в нацистском духе » [...] (Trad. Ils ont recours à l'intimidation par des nettoyages ethniques dans un esprit nazi).

Ces représentations sont intéressantes à étudier sur différents niveaux d'analyse :

- **Lexical :**Nous y repérons l'emploi d'un lexique péjoratif, religieux ou familial (« наш дом » notre maison, « наши братья и сёстры » nos frères et sœurs, « хаят нашу Родину-мать » ils blâment notre Mère-Patrie, « в нацистском духе » dans un esprit nazi) ;
- **Lexico-sémantique :** Des formes verbales permettant de modaliser le propos, comme « искусственные геополитические конструкции » (Trad. constructions géopolitiques artificielles), ou « лепят образ врага » (Trad. ils façonnent l'image de l'ennemi), « СМИ, подчиненные англо-саксонскому миру » (Trad. les médias sont soumis au monde anglo-saxon) ; ainsi que la convocation de formules intertextuelles : « На протяжении веков западные элиты привыкли наполнять свои животы человеческим мясом и карманы деньгами. Но они должны понять, что бал вампиров подходит к концу » (Trad. Depuis des siècles, les élites occidentales ont pris l'habitude de se remplir le ventre de chair humaine et les poches d'argent. Mais elles doivent comprendre que leur bal des vampires est en train de se terminer). Cette formule (bal des vampires) a largement été reprise du discours de Poutine lors de l'interview à Tucker Carlson dans les discours francophones et russophones.
- **Pragmatique et syntaxique :** Formes d'hétérogénéité énonciative, comme modalisateurs « pseudo-… » / « soi-disant… **»**, etc., anaphores à valeur axiologique, de constructions antithétiques, de questions rhétoriques exclamatives, argumentatives (comme dans 18) et conflictuelles **(***rhetorical opposing questions* - Gruber, 2001, comme dans 19) **:**

(16) « Для оправдания своих пагубных действий они придумывали псевдоморальные обоснования » (Trad. Pour justifier leurs actions néfastes, ils « inventaient des justifications pseudo-morales [...] ») ;
(17) Самонадеянность наших так называемых партнёров (Trad. L'arrogance de nos soi-disant partenaires [...]) ;
(18) « Когда ваш покорный слуга высказал предположение: а может нам и в НАТО вступить? Но нет, в НАТО такая страна не нужна. [...] В чём проблема? Видимо, проблема в геополитических интересах и надменном отношении к другим.» (Trad. Quand votre humble serviteur a émis l'hypothèse : peut-être devrions-nous rejoindre l'OTAN ? Mais non, un tel pays n'est pas nécessaire à l'OTAN. [...] Quel est le problème ? Apparemment, le problème réside dans les intérêts géopolitiques et l'attitude hautaine envers les autres ») ;
(19) « Всё время звучит : вы должны, вы обязаны, мы вас серьёзно предупреждаем... Вы кто такие вообще? Какое вы имеете право кого-то предупреждать? » (Trad. Tout le temps, on entend : vous devez, vous êtes obligés, nous vous avertissons sérieusement... Qui êtes-vous donc ? Quel droit avez-vous de prévenir qui que ce soit ? »).

- **Rhétorique :** Recours à l'argumentation fallacieuse, aux stratégies de discréditation, de l'homme de paille, à la rhétorique de l'évidence (force de choses), à l'analogie historique (par précédent), ainsi qu'aux effets de pathos (à visée de dramatisation et de déshumanisation) et de l'ethos (mise en scène d'un héros moralement supérieur). Par manque de place, nous ne développerons pas ce point ici.

Le corpus B, bilingue et multimodal, est composé de *commentaires discursifs conversationnels* (Paveau, 2017), postés en réaction à 300 articles et tweets évoquant, directement ou indirectement, le conflit russo-ukrainien (CRU), collectés sur X (anciennement Twitter). La sélection des publications repose sur l'analyse des profils de médias ainsi que de *leaders d'opinion* à forte audience, exprimant des positions contrastées et souvent tranchées sur l'invasion russe. Le choix des contenus retenus reflète les grandes étapes du CRU ayant fortement marqué l'opinion publique : l'annonce de l' « opération militaire spéciale », les premières offensives russes, les négociations à Istanbul, les événements de Boutcha, les contre-offensives ukrainiennes, les tensions diplomatiques, ainsi que les discours irréconciliables sur la Crimée. Des événements connexes contribuant à la polarisation des opinions ont également été pris en compte : l'interview de Vladimir Poutine par Tucker Carlson (en février 2024), la mort d'Alexeï Navalny (un des leaders de l'opposition russe), le discours de Poutine à la Fondation Valdaï sur le monde multipolaire, et celui de Volodymyr Zelensky à Davos sur la reconstruction de l'Ukraine comme projet européen de liberté. Sous ces publications, les échanges entre internautes abondent.

Les données de ce corpus sont codifiées et anonymisées, enregistrées sous forme de captures d'écran et répertoriées dans un tableau récapitulatif dont l'extrait est présenté ci-dessous à titre illustratif :

Code, réseau & communauté ou page (facultatif)	Date de publication	Date de consultation	Nombre de commentaires (au total)	Nombres de commentaires exploitables
GI_8 (X)	05/11/23	06/11/23	829	692

table 1. : Tableau 1 : Exemple de l'organisation des méta-données du corpus B

figure . 1 Figure 1 : Extraits du corpus B

Nous montrerons que certaines thèses élaborées dans les discours politico-médiatique officiels trouvent un écho contradictoire dans les discours socionumériques, russophones et francophones, et sont réappropriées par les locuteurs ordinaires. Ces extraits seront également commentés. Voici quelques exemples :

Notre analyse de la construction discursive de la figure de l'ennemi s'inscrira dans une approche rhétorico-interactionnelle. Nous mobiliserons les catégories classiques de la rhétorique — ethos, logos, pathos — tout en tirant parti de l'exploitation de deux corpus, ce qui nous permettra d'évaluer à la fois la portée illocutoire et perlocutoire des discours (comme les extraits de la figure 1 peuvent en témoigner, les reprises interdiscursives seront particulièrement intéressantes à examiner).

Bibliographie :

Bailly, N.L., & Moïse, C. et al. (2023). *Discours de haine et de radicalisation. Les notions clés*. Lyon : ENS Editions.

Billion, D., & Ventura, Ch. (2023). *Désoccidentalisation. Repenser l'ordre du monde*. Marseille : Agone. `

Charaudeau,P. (2006). *Discours journalistique et positionnements énonciatifs. Frontières et dérives*. Semen. Numéro 22. P.1-13.

Flaxman, Seth, Goel, Sharad, Rao, Jennifer M. (2016) "*Filter bubbles, echo chambers, and online news consumption."* [In:] *Public Opinion Quarterly*, 80(S1), 298–320.

Gruber, Helmut (2001) "Questions and strategic orientation in verbal conflict sequences." [In:] *Journal of Pragmatics*, 33; 1815–1857.

Huszár, Ferenz, Ribeiro, Angelo, Silva, Carlos (2021) "Algorithmic amplification of political content on Twitter." [In:] *Proceedings of the National Academy of Sciences*, 118(42), e2025334119.

Krieg-Planque, A. (2003). *Purification ethnique »: Une formule et son histoire*. Paris : CNRS Éditions.

Micheli, R. (2014). *Les émotions dans les discours : Modèle d'analyse, perspectives empiriques.* Paris : De Boeck.

Moirand, S.*(2007). Les discours de la presse quotidienne. Observer, analyser, comprendre, Paris, PUF (Linguistique nouvelle).*

Moïse, C. (2012). *Violence verbale, fulgurances au quotidien. Comprendre et agir.* Montpellier : Crdp.

Paveau, M.-A. (2017) *L'analyse du discours numérique. Dictionnaire des formes et des pratiques.* Paris : Hermann Editeurs.

Plantin, Ch. (2016) *Dictionnaire de l'argumentation. Une introduction aux études d'argumentation*, Lyon, ENS Éditions.

Semo, M. (2021). Fini les illusions de l'après-guerre froide : le retour de "l'ennemi", nouvelle réalité pour la France. *Le Monde.* Consulté le 30/11/2023 sur : https://www.lemonde.fr/politique/ar-ticle/2021/04/16/fini-les-illusions-de-l-apres-guerre-froide-le-retour-de-l-ennemi-nou-velle-realite-pour-la-fran-ce_6076978_823448.html].

Semo, M. Le "Sud global", cet ensemble hétérogène de pays non alignés. *Le Monde,* 26 octobre 2022. Accessed on April 12, 2025 : https://www.lemonde.fr/idees/article/2022/10/26/le-sud-global-cet-ensemble-heterogene-de-pays-non-alignes_6147333_3232.html.

Pages Web :

Ageeva, V. (2022). *Le soutien de la société russe à la guerre en Ukraine : sur les traces de l'homo sovieticus*. SciencesPO, centres de recherches internationales. URL : https://www.sciencespo.fr/ceri/fr/content/dossiersduceri/le-soutien-de-la-societe-russe-la-guerre-en-ukraine-sur-les-traces-de-l-homo-sovieticus (accessed on March 13, 2025).

Cros, E. (2003), La Sociocritique. Paris : L'Harmattan, extrait repéré à https://www.sociocritique.fr/?Redefinir-la-notion-d-ideologeme (accessed on March 25, 2025).

Simakova, M. (2023). Russie : la morale de Vladimir Poutine. Le Grand Continent (trad. par Guillaume Lancereau). URL : https://legrandcontinent.eu/fr/2023/12/25/russie-la-morale-de-vladimir-poutine/ (accessed on April 2, 2025).

GREMO-LING : Constituer un corpus de discours de patients présentant des lésions cérébrales acquises pour l'annotation et l'évaluation de leur expression émotionnelle

Salomé Klein [1], Amalia Todirascu,[1] et Hélène Vassiliadou [1]
[1] Laboratoire Linguistique Langues Parole (LiLPa, UR 1339), Université de Strasbourg (France) & Institut Thématique Interdisciplinaire LiRiC (Langage, Inclusion, Remédiation, Interculturalité et Communication)
salklein@unistra.fr, todiras@unistra.fr, vassili@unistra.fr

Introduction

Ce travail s'insère dans le cadre général de l'étude interdisciplinaire et participative réunissant des cliniciens et des linguistes « Regulating Emotions and Behaviors After Brain Injury » (NCT 05 39 34 92) et plus particulièrement dans le projet Idex Recherche Exploratoire GREMO-LING de l'Université de Strasbourg. Le projet vise à évaluer les bénéfices de la thérapie comportementale dialectique (TCD, Linehan, 2017) chez des patients atteints d'une lésion cérébrale acquise (LCA) et présentant une dysrégulation émotionnelle qui impacte négativement leur qualité de vie (Kuppelin et al., 2024). Dans ce contexte, nous tirons parti d'indices linguistiques présents dans leur discours, afin de fournir des mesures autres que celles utilisées par les médecins pour évaluer l'efficacité de la TCD. Notre travail constitue ainsi une perspective applicative dans le domaine de la santé de l'analyse d'interactions orales. Notre corpus d'entretiens de patients est un témoin original et précieux de ce type de discours atypique, que nous exploitons dans une perspective interdisciplinaire pour l'évaluation de la thérapie.

Cette communication a pour objectif d'exposer la méthodologie d'acquisition et d'annotation des données issues du discours de patients. Nous présenterons dans un premier temps la chaîne de traitement pour l'acquisition et la transcription du corpus. Nous mettrons tout particulièrement l'accent sur les problématiques rencontrées lors de la constitution du corpus, ce dernier devant répondre aux besoins des différents membres du projet. Dans un second temps, nous décrirons le guide d'annotation en expressions émotionnelles créé spécifiquement pour annoter manuellement le corpus tout en tenant compte de ses spécificités dues à l'oral. Nous proposons dans un troisième temps une analyse pilote du discours de deux patientes.

Corpus et méthodologie

Corpus

Le corpus est constitué de transcriptions d'enregistrements de patients. Chaque patient, au cours de son inclusion dans le protocole, est enregistré 3 fois : le premier enregistrement, T0, correspond à la baseline ou phase de contrôle. Le patient ne suit pas la thérapie TCD à ce moment-là mais bénéficie d'un suivi médico-social : psychologue, psychiatre, assistant social,

orthophoniste… Le deuxième enregistrement, T1, intervient 5 mois après directement avant le début de la thérapie. La comparaison entre l'enregistrement contrôle et l'enregistrement T1 effectué juste avant le protocole permet de soustraire aux effets de la thérapie la progression naturelle du patient. Durant les 5 mois après le second enregistrement, le patient suit la thérapie comportementale dialectique adaptée pour patients LCA (Goetsch, 2021; Kuppelin, 2024). Celle-ci se compose de séances de thérapie individuelle, de 3h hebdomadaires de séance de groupe d'apprentissage de compétences de régulation émotionnelle, ainsi que de la possibilité de faire appel à un coach téléphonique. A l'issue des 5 mois de thérapie, le patient est enregistré une troisième et dernière fois.

Les enregistrements sont effectués par les doctorants et stagiaires en orthophonie du projet et durent entre 30 minutes et 1 heure 30. Une attention particulière a été portée à la qualité sonore de l'enregistrement, avec une distance fixe du micro-cravate au locuteur, afin que des analyses prosodiques fines puissent être effectuées.

Dans les trois enregistrements, l'entretien est semi-dirigé et suit la même trame de questions. Il se fait en deux phases : la première partie aborde des questions autour de souvenirs associés à des émotions et de la personnalité et comporte des questions comme « racontez-moi un souvenir précis des 5 derniers mois où vous avez vécu une situation de détresse émotionnelle, une situation où vous étiez mal ». La seconde partie se concentre sur les thématiques du libre-arbitre et la capacité du patient à faire des choix. Dans le cadre de cette communication nous utiliserons exclusivement les entretiens issus de la première partie. Il est à noter que les relances lors des entretiens sont faites de manière minimale, en répétant les derniers mots du patient uniquement, afin de ne pas introduire de biais d'interprétation ou de changement de sujet.

Les enregistrements contiennent des données médicales sensibles et ils sont donc ensuite déposés sur un serveur sécurisé puis transcrits automatiquement en local par un outil utilisant le modèle de reconnaissance Whisper (Radford et al., 2023) et sa bibliothèque Whisper X (Bain et al., 2023). Cette première étape de transcription automatique permet de récupérer les tours de parole, une segmentation en phrase, ainsi que de maintenir l'alignement au signal sonore avec la récupération des bornes temporelles des mots prononcés. Les transcriptions sont ensuite corrigées à la main en suivant un guide de correction. Cela permet d'ajouter les segments manquants, anonymiser les noms propres et ajouter des informations de disfluence comme les répétitions, les révisions et les hésitations. Cette étape de correction permet de créer une version enrichie du corpus qui puisse répondre aux différentes questions de recherche des membres du projet.

L'extrait du corpus que nous présentons dans le cadre de cette présentation se compose de 6 transcriptions correspondant aux 3 temps d'enregistrement T0, T1 et T2 de deux patientes ayant bien répondu à la thérapie (score de 1 ou 2 sur 7 à l'échelle Clinical Global Impression of Improvement CGI-i, Busner & Targum, 2007). Le corpus fait un total de 2h51min07s et 20675 tokens répartis dans les 1246 phrases du discours des patientes. Ces six transcriptions ont ensuite été annotées manuellement en expressions émotionnelles.

Méthodologie

Le schéma d'annotation (Klein et al., 2024 ; Klein, 2025) prévoit une annotation en deux niveaux afin de s'adapter à la structure de l'oral, notamment dans le cadre du discours pathologique (variations de fluence, répétitions et révisions importantes). Le niveau plus général est celui de la phrase. Il permet d'encoder des informations d'émotionnalité (oui/non), de polarité (positif, négatif, mixte, incertain), d'intensité (intense/non-intense) et de catégorie émotionnelle (parmi celles enseignées dans le cadre de la thérapie et qui correspondent pour la

plupart aux émotions dites primaires en psychologie ; Ekman, (1971) : colère, joie, honte, dégoût/lassitude, tristesse, peur, amour, jalousie/convoitise) de la phrase en cours. La segmentation en phrase correspond à la délimitation introduite au niveau de la transcription automatique.

Le second niveau, plus précis, de l'expression émotionnelle, permet d'annoter les segments responsables de l'interprétation de la phrase comme émotionnelle. Une annotation des expressions émotionnelles en « mode d'expression » permet de renseigner si le segment annoté exprime une émotion de manière directe en ayant recours à du vocabulaire lexicalement codé (émotion Désignée : *frustration*, *être au fond du trou*) ou bien si cette expression est plus indirecte et passe par le contexte déclencheur (émotion Suggérée : *un conflit*, *perdre un être cher*) ou les conséquences physiologiques et comportementales d'une émotion (émotion Manifestée : *pleurer*, *serrer les dents*).

Résultats

L'annotation finale se compose de 1242 phrases annotées et corrigées, réparties comme suit :

Patient.e	nombre de phrases	nombre de phrases émotionnelles	pourcentage de phrases émo/phrases	nombre d'expressions émotionnelles	expressions Désignées	expressions Suggérées	expressions Manifestées	expr Désignées /total expr	expr Suggérées /total expr	expr Manifestées /total expr
Lia T0	218	146	66,97%	227	82	98	42	36,12%	43,17%	18,50%
Lia T1	163	89	54,60%	133	42	71	16	31,58%	53,38%	12,03%
Lia T2	339	190	56,05%	288	101	124	55	35,07%	43,06%	19,10%
Eva T0	129	68	52,71%	97	28	59	9	28,87%	60,82%	9,28%
Eva T1	239	104	43,51%	136	48	69	14	35,29%	50,74%	10,29%
Eva T2	154	83	53,90%	113	37	50	20	32,74%	44,25%	17,70%

table 1. : Nombre de phrases et d'expressions émotionnelles pour 2 patientes aux 3 temps de la thérapie

Ces annotations analysées, elles permettent de tirer un certain nombre de conclusions sur l'évolution des patients après la thérapie TCD. Il est à noter que les patients sont leurs propres contrôles, en comparant l'évolution de la période T0-T1 à l'évolution de la période T1-T2. Par exemple, le discours de la patiente Eva voit la proportion de ses expressions émotionnelles Manifestées, associées aux comportements liés aux émotions, augmenter (de 9,28% à 17,70%) tandis que la part d'expression émotionnelle Suggérée baisse (de 44,25/ à 60,82%). Cela se manifeste dans des phrases comme « Je réfléchis souvent... à enfin comment dire, pas m'emporter tout de suite, j'essaye. » (T2, phrase n°212) ou encore « Quand ça va pas trop […] alors j'observe souvent la nature, ça m'apaise. » (T2, n°207). La patiente semble être moins concentrée sur les contextes déclencheurs d'émotion et les stimuli et être plus en phase avec ses ressentis corporels, qu'elle a pu réapprendre à maîtriser grâce aux enseignements de la thérapie.

En somme, nous nous focaliserons sur la méthode d'acquisition et les caractéristiques d'un corpus constitué de discours de patients aux différents temps de leur suivi thérapeutique. Nous proposerons une analyse pilote de marqueurs linguistiques d'expression émotionnelle visant à mettre en valeur l'apprentissage des compétences de régulation émotionnelle issue de la thérapie TCD. L'annotation manuelle du corpus permet de mettre en valeur des points d'intérêt ainsi que faire des analyses quantitatives menant à la comparaison de l'évolution du patient aux 3 temps de son processus thérapeutique. Ce corpus se distingue par son originalité (discours oral et atypique), le secret médical compliquant l'acquisition de ce type de données. Les annotations ainsi que le guide d'annotation seront rendues disponibles prochainement, mais de par le secret médical, le corpus ne pourra pas être rendu disponible dans sa totalité. Le discours de patient reste enfin un moyen privilégié d'accès à son vécu et son évolution dans un cadre

thérapeutique. Le type d'étude que nous proposons s'avère nécessaire et ouvre des perspectives intéressantes dans le domaine de l'analyse des émotions et celui du traitement automatique de la langue.

Références bibliographiques

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). *WhisperX : Time-Accurate Speech Transcription of Long-Form Audio* (arXiv:2303.00747).

Busner, J., & Targum, S. D. (2007). The Clinical Global Impressions Scale : Applying

a Research Tool in Clinical Practice. *Psychiatry (Edgmont)*, *4*(7), 28-37.

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion.

Nebraska Symposium on Motivation, *19*, 207-283.

Goetsch, A. (2021). *Faisabilité d'une prise en charge par thérapie comportementale et dialectique de la dysrégulation émotionnelle chez des patients avec lésions cérébrale acquise en phase chronique et de son évaluation par auto-questionnaire.* [Thèse de doctorat, Université de Strasbourg].

Klein, S. (2025). Proposition d'un cadre méthodologique pour l'étude de l'expression linguistique des émotions. In A. Kananovich, R. Y. Belem, C. Lacassain, F. Marsac, B. Vaxelaire, & G. Kleiber (Éds.), *Regards sur la perception : De l'expérience au linguistique* (p. 53-72). Editions du CIPA.

Klein, S., Todirascu, A., Vassiliadou, H., Kuppelin, M., Becart, J., Briand, T., Coridon, C., Gehrard-Krait, F., Laroche, J., Ulrich, J., & Krasny-Pacini, A. (2024, mai). Annotating Emotions in Acquired Brain Injury Patients' Narratives. *Proceedings of CL4Health @ LREC-Coling 2024*. Patient-Oriented Language Processing, Turin, Italie. https://bionlp.nlm.nih.gov/cl4health2024/

Kuppelin, M. (2024). *La thérapie comportementale dialectique pour des personnes présentant une lésion cérébrale acquise : Adaptations, évaluations et caractérisation de la dysrégulation émotionnelle* [Thèse de doctorat, Université de Strasbourg].

Kuppelin, M., Goetsch, A., Choisel, R., Isner-Horobeti, M.-E., Goetsch, T., & Krasny-Pacini,

A. (2024). An exploratory study of dialectical behaviour therapy for emotional dysregulation and challenging behaviours after acquired brain injury. *NeuroRehabilitation*, *55*(1), 77-94.

Linehan, M. (2017). *Manuel d'entraînement aux compétences TCD* (N. Perroud, R. Nicastro, & P. Prada, Trad.; 2e éd). Médecine et hygiène.

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning*, 28492-28518.

Identifier les effets des reformulations sur l'utilisation de structures syntaxiques complexes : apports des analyses séquentielles d'interaction

Capucine Saulpic [1,2]
[1] Laboratoire Savoirs, Textes, Langage UMR8163, Université de Lille
[2] EA 7345 CLESTHIA, Université Sorbonne Nouvelle
capucine.saulpic@univ-lille.fr

Introduction

Pour étudier le développement langagier des enfants, de nombreux travaux se sont appuyés sur des corpus d'interactions en situations naturelles et ont montré le rôle du langage adressé à l'enfant (LAE) dans ce processus (voir Rowe & Snow, 2020 pour une synthèse). Plusieurs travaux ont mis en avant le rôle de la fréquence et de la quantité de LAE sur le développement de certains aspects langagiers (Ambridge et al., 2015). D'autres ont montré qu'engager des enfants dans des interactions contingentes joue également un rôle crucial (Masek et al., 2021). Dans ce contexte, différentes études ont examiné les caractéristiques des interactions dans l'environnement familial et ont souligné le rôle des reformulations par les adultes des énoncés des enfants pour leur développement phonologique, morphologique et lexical (Clark, 2020; Veneziano, 2014). Dans ces reformulations, l'adulte prend appui sur les verbalisations de l'enfant, lui fournissant ainsi l'expérience d'autres fonctionnements langagiers tout en répondant à ses hypothèses sur le fonctionnement du langage (de Weck, 2006). En contexte scolaire, des phénomènes similaires ont été observés (Farrow et al., 2020). Par ailleurs, les études sur l'émergence de la syntaxe ont mis en avant l'influence de la fréquence d'utilisation de certaines constructions par les adultes (Diessel, 2004) et du genre de discours (Canut et al., 2023). D'autres ont identifié les reformulations comme des pratiques soutenant l'émergence de la syntaxe complexe, montrant ainsi l'importance de l'étude d'interactions pour déterminer l'influence de ces pratiques sur le développement syntaxique (Canut et al., 2013). Néanmoins, ces travaux sont majoritairement qualitatifs et reposent sur des corpus restreints, ce qui ne permet pas de généraliser les résultats sur les processus interactionnels favorisant l'émergence de la syntaxe complexe après 3 ans.

Notre intérêt se porte sur l'acquisition des structures syntaxiques complexes car les enfants vulnérables sur le plan langagier possèdent généralement de plus faibles compétences dans ce domaine (Arndt & Schuele, 2013). Ces structures permettent d'organiser le discours, d'exprimer des relations logiques et d'évoquer le passé ou le futur, sans recourir au contexte d'énonciation (Canut & Vertalier, 2014). Dès lors, les enfants dits vulnérables sont plus susceptibles de présenter des difficultés pour produire des discours décontextualisés et accéder ensuite à la littératie (Conica et al., 2023; Gordon-Pershey, 2022). Pour pallier ces inégalités, plusieurs dispositifs d'aide ont été créés (Kern & Fekete, 2019). Toutefois, leur efficacité dépend de facteurs variés comme l'intensité du dispositif, l'expérience des participant·es ou encore les pratiques langagières employées (Markussen-Brown et al., 2017). Pour ce qui est des pratiques langagières, une revue systématique a corroboré le rôle des reformulations mais dans

certaines conditions seulement. Elles ne doivent pas être trop nombreuses et cibler certains éléments linguistiques (Cleave et al., 2015). Toutefois, les processus qui permettraient d'expliquer les acquisitions syntaxiques des enfants dans de tels dispositifs sont peu documentés.

Dans le cadre de ces journées d'étude, nous présenterons des analyses quantitatives et qualitatives réalisées sur un corpus d'interactions dyadiques recueillies dans le cadre d'un dispositif d'aide. Les premières consistent en des analyses séquentielles d'interaction réalisées avec le logiciel GSEQ (Bakeman & Quera, 2011). Elles permettent d'étudier l'effet immédiat de différentes pratiques langagières, dont les reformulations, sur l'utilisation de structures syntaxiques diversifiées par les enfants. A partir de l'identification de ces séquences adjacentes favorisant (ou non) l'utilisation de ces structures, nous avons mené des analyses qualitatives qui ont permis d'observer la mise en place de processus interactionnels et la complexité des phénomènes en jeu.

Corpus et méthodologie

Corpus

Notre étude s'appuie sur des données collectées au sein d'un dispositif d'aide dont l'objectif est d'accompagner les enfants dans leur développement syntaxique. Il s'agit des ateliers de langage « Je lis, on raconte » mis en place par l'Association de Formation et de Recherche sur le langage[1]. Ils ont lieu deux à trois fois par semaine sur le temps périscolaire et sont proposés à des enfants de Moyenne et Grande Section de maternelle qui, selon leurs enseignant·es, présentent des vulnérabilités langagières. Chaque enfant bénéficie de 15 à 20 minutes d'échanges individualisés autour de livres conçus pour les ateliers. Le choix de cette activité est lié aux avantages qu'elle présente pour le développement de du langage, notamment syntaxique puisqu'elle permet d'expérimenter des structures syntaxiques diversifiées. Lorsque l'enfant raconte, l'adulte doit l'aider à construire son récit et lui proposer des reformulations pour l'amener à allonger ses verbalisations et/ou à les complexifier. Pour ce faire, les intervenant·es bénéficient d'une formation sur les effets du LAE.

Méthodologie

53 enfants ont participé à l'étude, 21 étaient scolarisé·es en moyenne section (MS) et 32 en grande section (GS). Chaque enfant a été enregistré·e deux fois pendant les ateliers : en début et en fin d'année. Le corpus analysé est donc constitué de 116 enregistrements. Nous avons examiné les reformulations (i.e. reprise linguistique de tout ou partie d'un tour de parole de l'enfant) et les offres (i.e. apport de nouveaux éléments langagiers indépendamment des verbalisations antérieures de l'enfant) de l'intervenant·e. Puisque l'objectif des ateliers est des conduire les enfants à produire des variantes décontextualisées, nous avons étudié leurs tours de parole des enfants selon deux aspects : la complexité syntaxique (répartie selon trois degrés : simple, multiple et complexe) et la complétude. En effet, les structures complexes permettent d'établir des relations de dépendance entre les événements et, pour les comprendre hors contexte, il est nécessaire qu'elles soient suffisantes à elles-mêmes, autrement dit, complètes.

Nous avons ensuite mené des analyses séquentielles pour déterminer dans quelle mesure ces pratiques influencent chacun de ses aspects dans les verbalisations subséquentes des enfants, et si cette influence diffère en fonction de leur âge. Ces analyses permettent d'évaluer les

1 Jusqu'en 2025, ces ateliers étaient nommés « Coup de Pouce Langage » https://www.asforel.fr/

dépendances potentielles entre des comportements qui se succèdent. Il s'agit de déterminer si la présence d'un comportement source augmente la probabilité que le comportement cible se produise (Bakeman & Quera, 2011). Nous avons combiné ces analyses quantitatives à des études de cas afin de décrire les pratiques langagières des intervenant·es et les processus interactionnels qui se mettent en place au-delà d'une séquence adjacente.

Résultats

En début d'année, après une offre, les enfants de MS tendent à produire des tours de parole incomplets et les enfants de GS des interjections ou des onomatopées seules (figures 1 et 2). Ainsi, les offres permettent la poursuite de l'activité mais ne favorisent pas la production de tours de parole multiples ou complexes. Au même moment, les reformulations favorisent la production de structures complètes par les enfants de MS et n'ont pas d'effet sur les productions des enfants de GS. Les études de cas réalisées indiquent qu'en début d'année, reformulations et offres ne sont pas toujours adaptées au niveau linguistique de l'enfant, comme dans l'extrait ci-dessous (table 1.). L'adulte propose des reformulations et des offres avec le même degré de complexité que les verbalisations de l'enfant, ce qui ne lui permet pas d'expérimenter des structures plus longues ou plus complexes.

	Tour de parole	Analyse
CHI8	c'est + c'est la dame	Simple complet
ADU9	oui c'est la m- la madame et qu'est-ce qu'elle fait ?	Reformulation (simple complète) + question ouverte
CHI9	elle a appelé les < enfants	Simple complet
ADU10	bien > du coup chaque enfant + prend + un + <gâ- {en suspens}	Offre (simple incomplète)
CHI10	un gâteau >	Simple incomplet

table 1. : Extrait d'interaction entre l'enfant NI01C02 (MS) et la facilitatrice lors du premier enregistrement

En fin d'année, aucune de ces deux pratiques n'a d'effet sur les structures syntaxiques utilisées par les enfants de MS. En revanche, avec les enfants de GS, les reformulations favorisent l'emploi de structures complexes (figure 3). Plus précisément, elles peuvent donner lieu à une co-construction de tours de parole plus complexes que la production source de l'enfant, comme dans l'extrait ci-dessous (table 2.).

	Tour de parole	Analyse
CHI26	quand en ouvrant sa bouche la maman de Mélanie trouve ça sent bon et [e] {= ?} bien brossé ses dents	Complexe incomplet
ADU27	c'est en sentant la bouche de Mélanie que la maman voit qu'elle a bien brossé ses dents + vas y	Reformulation (complexe complète) + sollicitation
CHI27	quand Mélanie quand en sentant sa bouche s- la maman elle dit ça sent bon + et Mélanie range sa brosse et le dentifrice sur le [blø] (= ?} + gobelet et sur + et i(l) place sur l'étagère	Complexe incomplet
ADU28	bien elle range le dentifrice et la brosse à dents dans le gobelet qu'elle pose sur l'étagère	Reformulation (complexe complète)
CHI28	Mélanie + met [e] {= ?} brosse à dents sur le gobelet et son + et son dentifrice sur le gobelet et en posant le + sur le étagère	Complexe complet

table 2. : Extrait d'interaction entre l'enfant EB01S91 (GS) et la facilitatrice lors du second enregistrement

Ces deux extraits montrent que d'autres phénomènes influencent l'utilisation de certaines structures syntaxiques par les enfants. Par exemple, l'offre de l'adulte dans le premier extrait

est laissée en suspens, ce qui aboutit à la production d'un tour de parole incomplet. Dans le second, la facilitatrice encourage explicitement l'enfant à reprendre ses propositions. Ainsi, diverses pratiques langagières peuvent contribuer différemment à l'élaboration syntaxique des productions de l'enfant. Si l'analyse séquentielle permet d'obtenir des conclusions généralisables sur les effets de ces pratiques, elle est indissociable d'une analyse qualitative fine pour comprendre comment elles impactent les verbalisations des enfants en interaction.

Annexes

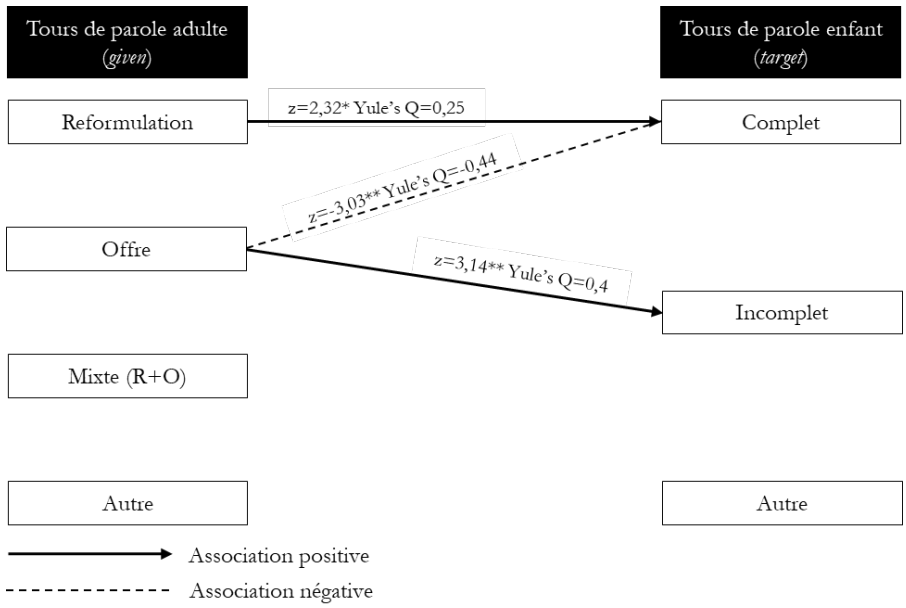


figure . 1 Associations entre les stratégies discursives contenues dans les tours de parole des facilitatrices (*given*) et le degré de complexité des tours de parole des enfants (*target*) en début de dispositif - *p<0,05, **p<0,01 (T1, MS)

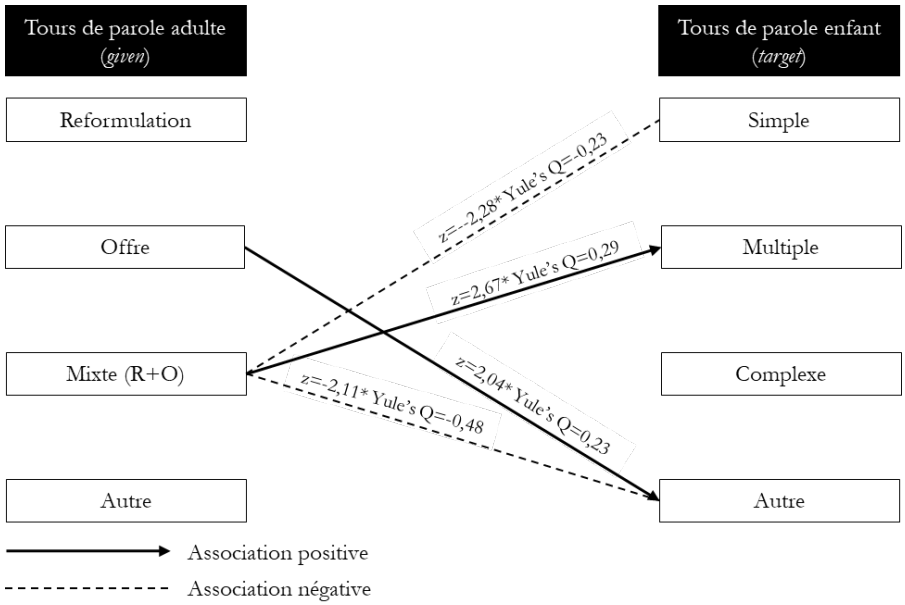


figure . 2 Associations entre les stratégies discursives contenues dans les tours de parole des facilitatrices (given) et le degré de complexité des tours de parole des enfants (target) en début de dispositif - *p<0,05, **p<0,01 (T1, GS)

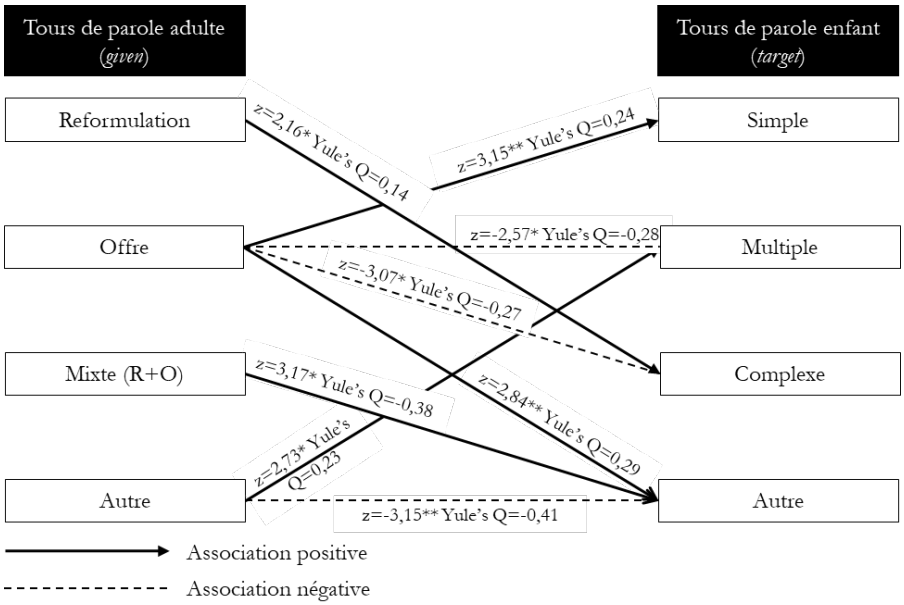


figure . 3 Associations entre les stratégies discursives contenues dans les tours de parole des facilitatrices (*given*) et le degré de complexité des tours de parole des enfants (*target*) en fin de dispositif - *p<0,05, **p<0,01 (GS, cohorte 2)

Références bibliographiques

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition*. *Journal of Child Language*, *42*(2), 239–273.

Arndt, K. B., & Schuele, C. M. (2013). Multiclausal Utterances Aren't Just for Big Kids: A Framework for Analysis of Complex Syntax Production in Spoken Language of Preschool- and Early School-Age Children. *Topics in Language Disorders*, *33*(2), 125–139.

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. Cambridge University Press.

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.

Canut, E., Jourdain, M., & Bocéréan, C. (2023). Developmental Markers of Pre-schoolers' Temporal and Causal Connectivity in Two Discourse Contexts: Data from the French Language. *Journal of Psycholinguistic Research*, *52*(6), 2373–2392.

Canut, E., & Vertalier, M. (2014). Les schèmes sémantico-syntaxiques créateurs dans le processus interactionnel d'acquisition. In N. Espinosa, M. Vertalier, & E. Canut, *Linguistique de l'acquisition du langage oral et écrit: Convergences entre les travaux fondateurs de Laurence Lentin et les problématiques actuelles* (pp. 85–116). L'Harmattan.

Clark, E. V. (2020). Conversational Repair and the Acquisition of Language. *Discourse Processes*, *57*(5–6), 441–459.

Cleave, P. L., Becker, S. D., Curran, M. K., Van Horne, A. J. O., & Fey, M. E. (2015). The Efficacy of Recasts in Language Intervention: A Systematic Review and Meta-Analysis. *American Journal of Speech-Language Pathology*, *24*(2), 237–255.

Conica, M., Nixon, E., & Quigley, J. (2023). Talk outside the box: Parents' decontextualized language during preschool years relates to child numeracy and literacy skills in middle childhood. *Journal of Experimental Child Psychology*, *236*.

de Weck, G. (2006). Les reprises dans les interactions adulte-enfant: Comparaison d'enfants dysphasiques et tout-venant. *La linguistique*, *Vol. 42*(2), 115–134.

Diessel, H. (2004). *The Acquisition of Complex Sentences*. Cambridge University Press.

Farrow, J., Wasik, B. A., & Hindman, A. H. (2020). Exploring the unique contributions of teachers' syntax to preschoolers' and kindergarteners' vocabulary learning. *Early Childhood Research Quarterly*, *51*, 178–190.

Gordon-Pershey, M. (2022). *Grammar and Syntax: Developing School-Age Children's Oral and Written Language Skills*. Plural Publishing.

Kern, S., & Fekete, G. (2019). De l'évaluation à l'intervention. In S. Kern (Ed.), *Le développement du langage chez l'enfant: Théorie, clinique, pratique* (De Boeck Supérieur, pp. 233–267).

Markussen-Brown, J., Juhl, C. B., Piasta, S. B., Bleses, D., Højen, A., & Justice, L. M. (2017). The effects of language- and literacy-focused professional development on early educators and children: A best-evidence meta-analysis. *Early Childhood Research Quarterly*, *38*, 97–115. https://doi.org/10.1016/j.ecresq.2016.07.002

Masek, L. R., McMillan, B. T. M., Paterson, S. J., Tamis-LeMonda, C. S., Golinkoff, R. M., & Hirsh-Pasek, K. (2021). Where language meets attention: How contingent interactions promote learning. *Developmental Review*, *60*, 100961.

Rowe, M. L., & Snow, C. E. (2020). Analyzing input quality along three dimensions: Interactive, linguistic, and conceptual. *Journal of Child Language*, *47*(1), 5–21.

Veneziano, E. (2014). Interactions langagières, échanges conversationnels et acquisition du langage. *Contraste*, *39*(1), 31.

Interagir avec un public virtuel sur YouTube : étude de la construction des interactions entre spécialistes du domaine de la mode et non-spécialistes à travers le corpus oral, numérique et multimodal VidEx

Agnès Ganet[1]
[1] Laboratoire ALTAE, Université Paris-Cité / Équipe LEADS, École Normale Supérieure Paris-Saclay
agnes.ganet@ens-paris-saclay.fr

Introduction

Les musées utilisent de plus en plus les nouvelles technologies, et notamment les réseaux sociaux, pour interagir avec leur public et attirer des visiteurs potentiels dans leurs murs (Capriotti et al 2016, Isik 2023). Nous assistons donc à une diversification et à une numérisation croissante des discours muséaux. Plus précisément, les ressources numériques sont notamment utilisées par les musées pour présenter les expositions temporaires qui y sont organisées.

On relève notamment une utilisation croissante par les musées de*YouTube*, site de partage de vidéos (Capriotti et al 2016). Ce site est de plus en plus perçu par les musées comme étant un outil important pour interagir avec leur public (*ibid*). Deuxième site le plus visité au monde et premier site le plus visité dans la catégorie « Arts et divertissements » d'après le site 'Top websites ranking' (2023), *YouTube* est très populaire (Godwin Jones 2007), et s'impose petit à petit comme une plateforme privilégiée pour partager des contenus numériques. En effet, *YouTube* se distingue par sa grande accessibilité (Shiryaeva et al 2019), car pouvant « atteindre virtuellement chaque coin (connecté) du monde » (Adami 2009). Ce site se distingue ainsi par « un mode de communication distinctif » (*ibid*), qui peut potentiellement aboutir à la création de communautés virtuelles, toutefois transitoires et désincarnées (Drasovean et Tagg 2015). Les « communautés virtuelles » sont définies par Rheingold (2013) comme « des agrégations sociales qui émergent du Net, lorsque suffisamment de personnes portent ces discussions suffisamment longtemps et avec assez de sentiments humains pour former des toiles de relations personnelles dans le cyberespace». L'élément fédérateur de ces communautés n'est autre que le langage (Drasovean et Tagg 2015), qui constitue un des piliers de *YouTube*, lequel devient dès lors une ressource précieuse et privilégiée pour constituer des corpus à la fois oraux, multimodaux et numériques.

Dans le cadre des vidéos postées par des musées sur *YouTube*, la communication est caractérisée, comme pour toutes les vidéos postées sur *YouTube,* par la séparation temporelle et spatiale des locuteurs et des auditeurs/spectateurs (Dynel 2014), mais aussi par leur méconnaissance mutuelle (*ibid*). En dépit de ces éléments, les locuteurs s'efforcent de créer un semblant d'interaction avec les internautes, ce qui a poussé Chovanec (2010) à parler de « quasi-interaction médiatisée ». Si l'étude de ces quasi-interactions peut être menée sur de nombreux

types de vidéos, la spécificité des vidéos issues du domaine muséal réside également dans leur caractère spécialisé et potentiellement vulgarisé, ce qui est susceptible d'avoir un impact notable sur la gestion de l'interaction.

Corpus

Le but et la raison d'être du corpus que nous avons constitué et que nous nous apprêtons à décrire est de caractériser les discours présents dans un certain type de vidéos issues du domaine muséal sur *YouTube.*

Cependant, à notre connaissance, un tel corpus n'existait pas, comme cela est souvent le cas pour les corpus spécialisés. De fait, nous avons dû créer notre propre corpus (Kübler 2014). L'avantage de cette démarche de création de corpus est que nous avons pu décidé de nos propres critères de sélection (*ibid*). Nous avons retenu cinq critères de sélection.

Le premier critère est disciplinaire : nous avons choisi d'inclure uniquement des vidéos portant sur des expositions de mode, telles que définies par Clark et De la Haye (2014). En effet, l'organisation d'expositions de mode temporaires dans les musées est une pratique non seulement récente (Riegels Melchior et Svensson 2014, Kurkdjian 2023), mais aussi très appréciée par le public (Clark et al 2014, Vrencoska 2015, Green et al 2019). Ces évènements occupent désormais une place importante voire omniprésente dans le paysage muséal (Vrencoska 2015, Mida 2015). Ces évènements sont également intéressants car ils sont souvent décrits comme ayant de multiples visées : les expositions de mode sont décrites dans la littérature et par les professionnels des musées comme des opportunités éducatives (Choi 2016, Green et al 2019), comme des opportunités récréatives en raison de leur caractère esthétique et spectaculaire (Mida 2015, Bedford 2015), mais aussi comme des opportunités économiques pour les musées (Kurkdjian 2024). Les expositions de mode se trouvent ainsi à la croisée de plusieurs visées, ce qui est susceptible d'avoir un impact sur les discours décrivant les expositions de mode.

Le second critère de sélection concerne la nature des évènements : nous avons choisi de travailler sur des expositions temporaires uniquement. En effet, il s'agit d'évènements éphémères (Green et al 2019), dont les enjeux sont différents des expositions permanentes (Isik 2023).

Ensuite, nous souhaitions étudier des discours spécifiquement muséologiques, afin d'étudier le caractère spécialisé de ces discours et la transmission de connaissances spécialisées par des spécialistes à des non-spécialistes. Aussi, nous avons uniquement inclus des vidéos partagées sur les chaînes *YouTube* des musées : ce cadre institutionnel constitue notre troisième critère de sélection.

Le quatrième critère concerne le cadre aréal de notre étude. Notre étude étant ancrée dans le domaine linguistique de l'anglais de spécialité, le domaine anglophone a été choisi. Plus précisément, nous avons choisi des musées britanniques et américains, le Royaume-Uni et les États-Unis occupant une place centrale dans le domaine des expositions de mode (De la Haye 2023). Aussi, 33 musées britanniques et 54 musées américains sont représentés dans le corpus.

Enfin, le cinquième critère de constitution du corpus est le cadre temporel de l'étude. Nous avons choisi 2008 comme borne de début du corpus et 2023 comme borne de fin. En effet, Mida (2015) a daté à 2008 le début de la « croissance exponentielle » des expositions de mode. En outre, le pic de la pandémie de Covid-19 en 2019 et 2020 a eu un impact significatif sur la communication muséale (Scatturo 2019) et notamment sur l'usage de *YouTube* par les musées (Suciu 2021). Les années de pandémie ayant été très spécifiques, nous avons choisi d'inclure

quelques années post-pandémie afin d'obtenir un échantillon plus représentatif de la communication muséale dans son ensemble. La perspective de notre étude demeure toutefois synchronique.

En définitive, nos critères peuvent être synthétisés comme suit : les vidéos doivent avoir été postées par des musées, et avoir pour thème des expositions de mode temporaires ayant eu lieu entre 2008 et 2023 au Royaume-Uni ou aux États-Unis. L'ensemble des ces vidéos a permis de constituer le corpus VidEx. Il a pour spécificité d'être un corpus oral et numérique en langue anglaise et est composé de 417 transcriptions de vidéos et de 234 639 mots. Le temps cumulé de vidéos s'élève à 29 heures et 15 minutes. Le corpus VidEx va donc permettre d'analyser et de caractériser le discours muséologique dans les vidéos de présentation d'expositions temporaires de mode, ce qui constitue une étude de cas des discours muséologiques numériques et oraux. L'outil numérique *YouTube* a été utile à plus d'un titre : au-delà de la disponibilité et de l'accessibilité à des discours spécialisés ou semi-spécialisés pour constituer notre corpus, nous avons pu utiliser l'outil « transcription » de *YouTube* pour compiler ce corpus (Coats 2023). Néanmoins, la qualité des transcriptions est loin d'être optimale et a donc requis des vérifications systématiques et minutieuses, aboutissant à la mise en place de stratégies de correction des erreurs de transcription.

Méthodologie

Afin de pouvoir étudier la mise en place d'interactions dans le corpus VidEx, et plus généralement afin de caractériser les discours qui le constituent, nous avons adopté une approche *corpus-driven* (Kübler 2014), soit une approche inductive et contextualisée (Koester, in O'Keeffe et McCarthy 2022). Le protocole mis en place est inspiré de la méthodologie de Partington (2006, 2010), et le logiciel *SketchEngine* a été utilisé.

Des corpus de référence ont été choisis afin de mener à bien cette analyse, la comparaison entre différents corpus permettant d'identifier les éléments caractéristiques et saillants du corpus étudié (Partington 2010). Les six corpus suivants ont été retenus : le corpus « TexModEx » et le corpus « Catalogues », constitués par nos soins, et qui sont des corpus de textes portant sur des expositions de mode ; le corpus « English Web 2021 », corpus de langue générale ; le corpus « British Academic Spoken English (BASE) », corpus de langue académique orale ; les corpus « Open American National Corpus (spoken) » et « British National Corpus (spoken part) », des corpus de langue orale et générale. Chacun de ces corpus présente des caractéristiques pertinentes pour être comparé au corpus VidEx, que ce soit l'oralité, le domaine spécialisé, ou encore la langue académique.

À terme, il s'agira de caractériser les discours issus du corpus VidEx, et notamment de déterminer l'ancrage spécialisé de ce corpus dans les domaines de la mode et de la muséologie, mais aussi les différentes visées caractéristiques de ce discours spécifique (promotion, instruction et information notamment).

Résultats

Si les interactions sont typiques de l'oral (Kerbat-Orecchioni 1998, Takahira 2014), et si les vidéos qui constituent le corpus VidEx relèvent bien du domaine oral (voire multimodal), elles sont pourtant dépourvues d'interactions réelles avec le public. Elles s'apparentent ainsi, selon le paradigme de Koch et Oesterreicher (1985) à du « discours de distance ». Si ce dernier présuppose un « caractère public, la domination des monologues, la faible implication émotionnelle, des thèmes fixés, la réflexion et donc un haut degré de planification » (Osthus 2018), les « discours de proximité » sont caractérisés par « des discours de familiarité entre

participants, le développement thématique libre, la coopération, la spontanéité et l'expressivité » (*ibid*).

En dépit de ces constatations, il semblerait que des stratégies linguistiques soient mises en place pour créer des interactions dans le corpus VidEx, et donc se rapproche d'un « discours de proximité » : la présence non négligeable des pronoms de première et de deuxième personne et les schémas lexico-grammaticaux associés (*you see*, *you can see*, *if you look at*) suggèrent la recherche d'une interaction. L'étude du corpus a également permis de mettre au jour la présence quasi-systématique dans les vidéos d'une étape de présentation du locuteur, et d'autres marques linguistiques telles que *hello*, *hi* ou encore *welcome*, qui ne sont pas sans rappeler les « rites d'interaction » typiques de l'oral (Goffman 1974). La présence statistiquement significative dans le corpus VidEx de marqueurs de discours tels que *you know*, *now*, ou encore *I mean*, dont le but est notamment de « rendre une conversation cohérente » (Chaume 2004), et qui sont donc caractérisés par leur vocation à construire une interaction, confirme l'hypothèse que ces vidéos sont le théâtre de « quasi-interactions médiatisées » (Chovanec 2010).

Toutefois, ces dernières ont non seulement pour spécificité d'être virtuelles (Adami 2009, Drasovean et Tagg 2015), mais aussi de s'opérer entre spécialistes d'un domaine et non spécialistes. Aussi, à cet aspect atypique de la production orale et des interactions qui y sont reliées s'ajoute la composante du spécialisé. En effet, la constitution du corpus VidEx a révélé son caractère spécialisé et vulgarisé. Des éléments statistiquement significatifs dans le corpus VidEx tels que *if we look at the* et *as you can see* permettent par exemple de guider et orienter l'attention du public vers un point d'intérêt, mais aussi de structurer et illustrer le propos dans un souci de clarté pédagogique.

Références bibliographiques

Adami, E. (2009). 'We/YouTube': exploring sign-making in video-interaction. *Visual Communication*, *8*(4), 379-399.

Bedford, Leslie. 2014. *The Art of Museum Exhibitions: How Story and Imagination Create Aesthetic Experiences*. https://ci.nii.ac.jp/ncid/BB16615843.

Capriotti, P., Carretón, C., & Castillo, A. (2016). Testing the level of interactivity of institutional websites: From museums 1.0 to museums 2.0. *International journal of information management*, *36*(1), 97-104.

Chaume, F. (2004). Discourse markers in audiovisual translating. *Meta*, *49*(4), 843-855.

Choi, K. H. (2016). Fashion criticism in museology-The Charles James retrospective. *Journal of the Korean Society of Clothing and Textiles*, *40*(3), 437-455.

Chovanec, J. (2010). Online discussion and interaction: The case of live text commentary. In *Cases on online discussion and interaction: Experiences and outcomes* (pp. 234-251). IGI Global Scientific Publishing.

Clark, J., & De La Haye, A. (2014). *Exhibiting fashion: Before and after 1971*. Yale University Press.

Coats, S. (2023). Dialect corpora from YouTube. *Language and linguistics in a complex world*.

Drasovean, A., & Tagg, C. (2015). Evaluative language and its solidarity-building role on TED. com: An appraisal and corpus analysis. *Language@ Internet*, *12*.

Dynel, M. (2014). Participation framework underlying YouTube interaction. *Journal of Pragmatics*, *73*, 37-52.

Godwin-Jones, R. (2007). Digital video update: YouTube, flash, high-definition.

Goffman, E. (1974). *Les rites d'interaction*. París: Ed. de minuit.

Green, D. N., Du Puis, J. L., Xepoleas, L. M., Hesselbein, C., Greder, K., Pietsch, V., ... & Estrada, J. G. (2021). Fashion exhibitions as scholarship: Evaluation criteria for peer review. *Clothing and Textiles Research Journal*, *39*(1), 71-86.

Işık, E. E. (2023). A corpus-based genre analysis of promotional-informational discourse in online painting exhibition overviews. *English for Specific Purposes*, *70*, 44-56.

Kerbrat-Orecchioni, C. (1998). La notion d'interaction en linguistique: Origines, apports, bilan. *Langue française*, 51-67.

Koch, P., & Oesterreicher, W. (1985). Sprache der Nähe—Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, *36*(1), 15-43.

Koester, A., in O'keeffe, A., & McCarthy, M. J. (2022). 'Of what is past, or passing, or to come 1': corpus linguistics, changes and challenges. In *The Routledge handbook of corpus linguistics* (pp. 1-9). Routledge.

Kübler, N. (2014). Mettre en œuvre la linguistique de corpus à l'université. Vers une compétence utile pour l'enseignement/apprentissage des langues?. *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle*, *11*(11-1).

Kurkdjian, S. (2024). Briller internationalement via la mise en culture de la mode. *Géopolitique de...*, *2*, 121-133.

Melchior, M. R., & Svensson, B. (Eds.). (2014). *Fashion and museums: Theory and practice*. A&C Black.

Mida, I. (2015). The Enchanting Spectacle of Fashion in the Museum. *Catwalk: The Journal of Fashion, Beauty, and Style*, *4*(2), 47-70.

Osthus, D. (2018). À la recherche du «locuteur ordinaire»: vers une catégorisation des métadiscours. *Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires*, (14), 18-32.7

Partington, A. (2006). Metaphors, motifs and similes across discourse types: Corpus-Assisted Discourse Studies (CADS) at work. *Trends in linguistics studies and monographs*, *171*, 267.

Partington, A. (2010). Modern diachronic corpus-assisted discourse studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora*, *5*(2), 83-108.

Rheingold H. (2000), The Virtual Community, homesteading on the electronic frontier, MIT press.

Scaturro, S. (2019). "Fashion as an Event: Conservation and Its Digital (Dis)Contents." *Critical Studies in Fashion and Beauty* 10 (1): 113–27. https://doi.org/10.1386/csfb.10.1.113_1.

Shiryaeva, T., Arakelova, A., Golubovskaya, E., & Mekeko, N. (2019). Shaping values with" YouTube freedoms": linguistic representation and axiological charge of the popular science IT-discourse. *Heliyon*, *5*(12).

Suciu, Marta-Christina, Gheorghe-Alexandru Stativă, and Mircea-Ovidiu Mitucă. 2021. "Creative and cultural sectors during the COVID-19 pandemic." In *Proceedings to the 4th International Conference on Economics and Social Sciences. Resilience and Economic Intelligence Through Digitalization and Big Data Analytics*, pp. 288-298. DOI: 10.2478/9788366675704-029

Takahira, Y. (2014). Discourse analysis of interpersonal features in ESL and JSL textbooks.

'Top Websites Ranking' (2023). Similarweb, consulté le 8/12/2023 <https://www.similarweb.com/top-websites/> et <https://www.similarweb.com/top-websites/>

Vrencoska, G. (2015). Museum Fashion Exhibitions: The fashion designer as an artist and new paradigms of communication with the audience. *New space in art and science*, *515*, 528.

J'arrive pas à faire des longs vocaux il faut *que je réfl~ respire entre les deux hum…* À propos des amorces de mots dans un corpus de vocaux

Anne Dister
UClouvain - Saint-Louis Bruxelles
anne.dister@uclouvain.be

Les disfluences

Les études sur la langue parlée ont permis de dégager des phénomènes propres à l'oral, qu'on regroupe souvent sous l'appellation générale de *disfluences*. On entend par là un certain nombre de traits liés à la production de la langue parlée, d'« achoppements » dans la linéarité de l'énoncé, de marques du discours en cours d'élaboration. Ces phénomènes sont inhérents aux productions orales, même si leur fréquence semble dépendante de la planification ou non de l'énoncé.

Au lieu de la disfluence, le déroulement linéaire est brisé et l'on assiste à un piétinement sur l'axe syntagmatique, qui a été modélisé par Shriberg (1994 : 7-9), à la suite notamment de Levelt (1983).

Les amorces

Dans cette communication, nous nous intéressons à l'une de ces marques de disfluence, les amorces de mots. Nous appelons *amorce* le phénomène langagier qui consiste en « une interruption de morphèmes en cours d'énonciation » (Pallaud 2002 : 79). Selon Blanche-Benveniste *et al.* (1990), l'amorce participe d'un phénomène d'anticipation, marqué par des allées et venues sur l'axe syntagmatique.

L'exemple suivant est un cas typique d'amorce. Le morphème interrompu – noté dans la transcription par un tilde collé directement à la droite de celui-ci, au lieu de l'interruption – est complété plus loin dans l'énoncé, où il est repris sous sa forme pleine :

> (1) il est parti ce matin très tôt et il revient pas avant vendr~ vendredi matin (<285_19>, F, 29, F)

Typologie et analyse des amorces

Dans cette communication, nous analysons 300 amorces issues du corpus *Les Vocaux* (Glikman et Fauth 2022).

Nous reprenons la typologie de Pallaud (2002), qui distingue les amorces complétées (cf. l'exemple 1 ci-dessus), les amorces corrigées (2) et les amorces abandonnées au profit d'une autre construction (3) :

(2) puis elle a répon~ elle a dit faut faire une danse de la pluie (<365_26>, A, 34,B)

(3) mais en même temps si je réussis sans heures d'auto-école je vais me dire ah ben c'est cool j'ai économi~ enfin après c'est pas pour l'argent que ça coute (<105_23>, F, 22, B)

Certaines amorces du corpus sont inclassables, d'autres participent d'une séquence de plusieurs amorces qui s'enchainent. La répartition selon le type d'amorce est la suivante :

Types d'amorces	Ventilation
complétées	47,6 % (143)
corrigées	19 % (57)
inachevées	18,6 % (56)
inclassables	3,6 % (11)
séquences	11 % (33)

table 1. : ventilation selon le type d'amorces

Si certaines amorces sont inclassables, c'est notamment parce que la séquence amorcée est souvent très brève (une seule consonne ou une seule voyelle), comme l'illustre le tableau suivant qui indique la longueur de l'amorce :

Longueur	n=
consonne	171
voyelle	20
1 syllabe	90
2 syllabes	16
3 syllabes	2
4 syllabes	1

table 2. : longueur de l'amorce

Contrairement à ce que l'on rencontre dans le phénomène de répétition (Dister 2007 ; Henry et Pallaud 2003 ; Pallaud et Henry 2004), les amorces concernent majoritairement les mots lexicaux :

Classe grammaticale	N=
nom	43
verbe conjugué	35
pronom sujet	29
verbe à l'infinitif	23
déterminant	18
adverbe	14
participe passé	14
pronom clitique	13
préposition	9
adjectif	8
adverbe composé	7
nom propre	3
conj. coordination	3

conj subordination	2
ponctuant	2
pronom relatif	2

table 3. : classe grammaticale des amorces

Nous comparons ensuite nos résultats avec ceux obtenus dans des conversations non planifiées. Nous pouvons conclure de notre étude que le phénomène disfluent de l'amorce de mots ne distingue pas les vocaux, énoncés généralement brefs qui obéissent à des contraintes propres, de données orales obtenues dans le cadre de longs entretiens semi-dirigés.

Références bibliographiques citées

Blanche-Benveniste, Cl., Bilger, M., Rouget, Chr., van den Eynde, K. (1990). *Le Français parlé. Études grammaticales*. Paris. CNRS Éditions.

Dister, A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales Valibel*. UClouvain.

Glikman J., C. Fauth (2022). Un nouvel accès à la parole spontanée : les vocaux. *34e Journées d'Études sur la Parole, JEP2022*, 154 162. ISCA. doi.org/10.21437/JEP.2022-17

Henry, S., Pallaud, B. (2003). Word fragments and repeats in spontaneous spoken French, R. Eklund (Éd.), *Proceedings of DISS'03. Disfluency in Spontaneous Speech Workshop,* (5-8 Septembre 2003, Göteborg University, Sweden), Gothenburg Papers in Theorical Linguistics 90, pp. 77-80.

Levelt, W. J.M. (1989). *Speaking : from intention to articulation.* Cambridge. MIT Press.

Pallaud, B. (2002). Les amorces de mots comme faits autonymiques en langage oral. *Recherches sur le français parlé 17*, Université de Provence, 79-101.

Pallaud, B., Henry, S. (2004). Amorces de mots et répétitions dans les énoncés oraux, *Recherches sur le français parlé 18.* pp. 201-229.

Shriberg, E. (1994). *Preliminaries to a Theory of Speech Disfluencies, Université de Berkeley*, Thèse non publiée.

L'analyse multimodale d'un corpus de discours de médiation scientifique sur YouTube : le projet Sciences en Herbe

Alexia Jingand[1], Clotilde George[1]
[1] Laboratoire ATILF, Université de Lorraine
alexia.jingand@univ-lorraine.fr, clotilde.george@univ-lorraine.fr

Introduction

Les enjeux environnementaux occupent une place importante dans diverses sphères discursives, qu'il s'agisse du discours politique (Fracchiolla, 2019), économique (Vignes, 2023), militant (Allouche, 2015) ou encore des discours protéiformes sur les réseaux sociaux (par exemple Lippert & Wagener, 2024). L'un des lieux privilégiés par la communication écologiste sur le terrain numérique correspond à la plateforme audiovisuelle YouTube, qui a déjà fait l'objet de plusieurs recherches en analyse du discours environnemental (Lartigue, 2020). Notre communication s'inscrit dans cette continuité et se concentre sur les défis rencontrés lors de la collecte et du traitement de données de médiation scientifique à portée écologique issues de YouTube. Plus précisément, nous reviendrons sur la constitution et l'analyse d'un corpus exploratoire réalisé dans le cadre du projet Sciences en Herbe, consacré à l'information scientifique et technique (IST) dans les discours de youtubeur·euses (Chevry Pébayle, 2021) et portant sur la tonte des espaces verts et ses enjeux de biodiversité, thématique encore peu exploitée dans les recherches linguistiques. Pourtant, il s'agit d'une question où se rencontrent des postures individuelles (tonte de jardin privé), collectives (normes esthétiques), politiques (fauchage tardif des espaces publics) au sein desquelles la recherche linguistique peut jouer un rôle descriptif essentiel à la compréhension des tensions cristallisées autour de tels enjeux.

Corpus et méthodologie

Corpus

Le corpus est constitué de six grands types de données :

- Les vidéos (30 vidéos pour le corpus exploratoire) : corpus_1a,
- Les transcriptions de l'oral (annotations relatives au matériau sonore) : corpus_1b,
- Les annotations relatives au matériau visuel des vidéos : corpus_1c,
- Les légendes textuelles accompagnant la vidéo (titres, textes de présentations, liens URL) : corpus_2,
- Les commentaires des vidéos (section commentaires de la plateforme) : corpus_3,
- Les annotations d'analyse (codage) : corpus_4.

Cette diversité complexifie le processus de collecte et de traitement, avec des enjeux techniques et plus globalement méthodologiques (éthiques, épistémologiques) liés à ces formats. La sélection des données du corpus_1a suit deux critères principaux. Tout d'abord, la production des vidéos se situe dans l'aire linguistique francophone, de sorte que la variation géographique constituera, à terme, un point de départ pour une approche contrastive sur les différents rapports à la tonte dans ces espaces. Par ailleurs, la durée des vidéos exclut le format très court (les «

shorts ») et les formats long (plus de quinze minutes) et très long (plus de 30 minutes), afin d'assurer la comparabilité des stratégies discursives.

Méthodologie

Les données récoltées feront l'objet d'une analyse du discours numérique (Paveau, 2017) dans la mesure où le discours audiovisuel n'est pas seul objet d'étude, puisque nous considérons comme données primaires l'environnement linguistico-numérique global avec les légendes et commentaires. L'approche présentée ici intègre une approche principalement multimodale et sociolinguistique. En effet, la nature multiple du corpus implique d'utiliser des outils adaptés à l'analyse multimodale (Kress & Van Leeuwen, 2001) et de conduire une réflexion sur leur articulation. Au centre de cette réflexion méthodologique se trouve le développement - dans l'environnement AVAA Toolkit [1] - d'une *pipeline* de traitements pour une constitution automatisée du corpus de transcription (corpus 1_b), telle qu'elle est réalisée dans le travail de Méli, Coats & Ballier (2023), à la différence qu'elle comprend l'interopérabilité avec ELAN[2] pour l'annotation multimodale manuelle (corpus_1c et 4). En outre, l'intégration de scripts de *scraping* opensource pour la collecte des légendes textuelles (corpus_2) et de commentaires (corpus_3), permet de constituer un unique corpus décliné en six sous-corpus interconnectés pour l'analyse. Celle-ci repose sur l'usage de plusieurs outils. Le logiciel TXM (Heiden, Magué & Pincemin, 2010) répond aux besoins de fouille et d'analyse textométrique, quantitative, des corpus 1b, 2 et 3 de production initiale écrite, tandis que le logiciel AVAA Toolkit répond aux besoins de fouille et d'analyse du corpus global par la production de collections de données audiovisuelles et textuelles combinées, et de visualisations adaptées à une analyse plurisémiotique, quantitative et qualitative. Ces outils convergeront vers une analyse du discours autour des stratégies discursives mises en œuvre par les youtubeur·euses et les réactions suscitées par ces vidéos. À partir de travaux sur les représentations sociales (Jodelet, 2003), l'exploration du corpus entend considérer à la fois les enjeux individuels, collectifs et politiques qui sous-tendent la question de la tonte des espaces verts en s'appuyant sur une lecture multimodale, plurisémiotique et textométrique permise par les outils de linguistique de corpus.

Résultats

Les résultats préliminaires suggèrent une intégration marginale de l'IST aux stratégies discursives des vidéos, où elle joue un rôle de légitimation pour les vidéastes. Les premières analyses font état d'une variété de profils de vidéastes, des profanes aux scientifiques vidéastes en passant par les « pro-ams »[3] qui ont recours aux arguments liés au bien-être et à la beauté, ou encore des municipalités qui utilisent le thème du fauchage tardif comme un argument économique. Néanmoins, les stratégies discursives multimodales ne semblent pas dépendre des enjeux de la vidéo ou de la source de diffusion : plans fixes sur les fleurs sauvages, métaphores conceptuelles sur les richesses de la nature, lexique de la vie de la nature, etc. se rencontrent dans l'ensemble du corpus sans distinction. Ces différentes stratégies argumentatives autour de la tonte ne semblent pas avoir le même effet sur le public dans la mesure où les lexiques

1 Audio and Video Annotations Analysis (George & George, 2025). URL : www.avaa-toolkit.org/

2 ELAN (Version 6.9). (2024). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. URL : https://archive.mpi.nl/tla/elan

3 Des amateurs dont la création de contenu correspond aux standards de vidéastes professionnels (Leadbeater & Miller, 2004).

développés dans le corpus de commentaires renvoient principalement à une caractérisation axiologique de la nature, insistant sur sa dimension esthétique. Les vidéos présentent relativement peu de vues, likes et commentaires, ce qui témoigne d'un intérêt encore limité pour les enjeux environnementaux qu'elles soulèvent. En outre, les commentaires reviennent peu sur l'IST. Ainsi, les données illustrent un rapport d'abord émotionnel à ces enjeux, rapport qu'il s'agirait d'aborder à partir d'une lecture anthropo-sociologique : les espaces verts mentionnés dans les vidéos sont surtout des jardins privés, sous la responsabilité individuelle, mais soumis au jugement collectif. Ces données confirment l'hypothèse d'un fractionnement entre représentations et pratiques à plusieurs échelles de la communauté. YouTube apparaît comme un lieu propice à la redéfinition en cours des normes dans certains espaces francophones autour des pratiques de tonte. Les stratégies discursives multimodales et la diffusion d'arguments sur le terrain numérique contribuent alors à une révision des représentations et normes écologiques.

Références bibliographiques

Allouche, A. (2015). L'argumentation dans la formation des groupes protestataires : du conflit d'aménagement au militantisme environnemental. *Argumentation et Analyse du discours*, 14, 1-15. https://doi.org/10.4000/aad.1929

Chevry Pébayle, E. (2021). Pratiques informationnelles des youtubeurs scientifiques au service de la médiation du savoir, *Communication* [En ligne], Vol. 38/2. URL : http://journals.openedition.org/communication/14808

Fracchiolla, B. (2019). Écologie et environnement : des mots aux discours. Mises en perspective historiques et discursives. *Mots. Les langages du politique*, 119, 14-31. https://doi.org/10.4000/mots.24230

Kress, G. R. et Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Arnold, Oxford University Press, London, New York.

Heiden, S., Magué, j.-P., & Pincemin, B. (2010). *TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement.* 10th International Conference on the Statistical Analysis of Textual Data.

Jodelet, D. (2003) [1989]. *Les Représentations sociales*. Presses Universitaires de France.

Lartigue, C. (2020). Crédible et objectif ou intime et émouvant : une analyse des stratégies discursives des vidéos de Youtube autour de l'environnement. *Questions de communication*, 38(2), 409-440. https://doi.org/10.4000/questionsdecommunication.24330

Leadbeater, C., & Miller, P. (2004). *The pro-am revolution: How enthusiasts are changing our society and economy*. Demos.

Lippert, E., & Wagener, A. (2024). Le mème comme technogenre. Analyse sémiodiscursive à partir d'un corpus de mèmes environnementaux. *Pratiques*, 203, 1-18. https://doi.org/10.4000/12ye8

Méli, A., Coats, S., Ballier, N. (2023). Methods for phonetic scraping of Youtube videos. In Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023), 244–249.

Paveau, M.-A. (2017). *L'Analyse du discours numérique.* Hermann.

Vignes, A. (2023). Réchauffement climatique, risqué d'accident : le discours dissonant des instances nucléaires françaises. Hamon, Y. & Paissa, P. (dir.). *Discours environnementaux : convergences et divergences.* Lingue d'Europa e del Mediterraneo.

L’évolution de marqueurs méta-discursifs dans une perspective longitudinale et interactive : le cas de ‘en vrai’.

Jeanne Sciberras
Centre de Linguistique Appliquée, Université de Neuchâtel
jeanne.sciberras@unine.ch

Introduction

Notre travail s’inscrit dans le champ de l’analyse conversationnelle et de la linguistique interactionnelle. Nous nous intéressons principalement aux expressions et marqueurs discursifs tels que : *genre*, *du coup*, *en vrai*, dans une perspective interactionniste et longitudinale. Souvent relégués à des énièmes tics de langage qui parasitent les discours et stigmatisent le “langage des jeunes”, ces expressions sont riches de sens et ont un rôle précis tant au niveau syntaxique, sémantique que pragmatique dans la cohésion du discours.

Notre recherche se base sur un corpus de conversations spontanées d’une trentaine d’heures enregistrées sous format audio et vidéo. Le corpus en lui-même (nommé Pauscaf) est constitué d’un focus synchronique de trois temporalités : le premier enregistrement s’est tenu en 2013, le second en 2019 et le troisième en 2024-25.

La perspective longitudinale que nous adoptons dans notre recherche, nous permet d’observer sur presque dix ans l’évolution de ces marqueurs. En complémentarité, la perspective interactionniste nous donne les outils pour analyser en usage le rôle, les fonctions et les actions qu’ont ces marqueurs. L’analyse conversationnelle (AC) suit une méthode inductive et a pour objectif de montrer que la conversation suit un ordre accompli par les participants dans l’interaction (Mondada, 2017). L’intérêt et l’accès aux interactions en contexte naturel a permis d’illustrer que l’interaction se soumet à des règles précises. Ces règles sont liées à l’agencement des tours de parole, montrant l’orientation des participants en fonction de ces tours, mais aussi la séquentialité des actions (Sacks et al. 1974). La linguistique interactionnelle de son côté porte son intérêt sur l’étude de la langue en interaction et les structures grammaticales du langage en usage. Ce focus analytique rend compte du fonctionnement des formes dans les discours (Pekarek Doehler, 2005 ; Mondada, 2008).

Dans le cadre de cette présentation nous aimerions concentrer notre attention sur la locution adverbiale ‘*en vrai’*. L’analyse de notre corpus sous un angle longitudinal nous as permis de suivre le développement de cette expression, et particulièrement dans le corpus Pauscaf2, (2019) et son expansion dans le corpus Pauscaf3, (2024-25). Notre focus analytique se porte sur les différents usages et fonctionnement qu’adopte *en vrai,* soit en fonction de son contexte d’emploi et de sa récurrence. Sommes-nous en train d’assister à l’essor de cette forme ? Quels sont ses panels d’emplois ? Pouvons-nous repérer un stade de pragmaticalisation ?

Corpus et méthodologie

Corpus

Corpus	**Pauscaf1**	**Pauscaf2**	**Pauscaf3**
Date d'enregistrement	2013	2019	2024-25
Nombre d'heures	9h	6h30	4h52[1]
Nombre de mots	110 883	77 988	47 920
Occurrences 'en vrai'	0	16	61
Nombres d'occurrence par 10'000 mots	0	2.05	12.72

table 1. : Corpus Pauscaf et apparition de 'en vrai'

Méthodologie

Notre recherche s'appuie sur ces trois sous-corpus qui ont été réalisés entre 2013 et 2025. Le corpus représente plusieurs paramètres importants à prendre en compte dans une analyse évolutive de la langue. L'objectif premier de la récolte des données était d'enregistrer des jeunes francophones lors de conversations spontanées. De ce fait, nous avons une certaine homogénéité diastratique car l'échantillon de la population enregistrée implique des étudiant.e.s universitaires entre 20 et 30 ans, rendant compte d'un groupe social précis. Le corpus a également une constance diaphasique, c'est-à-dire lié au niveau de langue, lui-même relatif au contexte de communication.

Les données recueillies sont transcrites avec une première transcription dite "normalisée" car elle contient uniquement le texte de la conversation. Cette transcription dépouillée de tous les signes de l'oral permet une quantification du corpus, notamment pour le nombre de mots, d'occurrences et de fréquences grâce à un logiciel de textométrie.

Par la suite, une deuxième transcription fine et détaillée est établie manuellement, qui suit les conventions de transcriptions usuelles en AC (Jefferson, 2004). Cette dernière constitue notre matériel principal pour l'analyse des données sous l'angle de l'analyse conversationnelle. Ces transcriptions rendent compte des détails interactifs comme les moments de chevauchement entre les interlocuteur.trice.s, les pauses, les caractères prosodiques (accélérations du débit, haussement de la voix, intonation montante ou descendante), ainsi que la séquentialité des tours de paroles. Une analyse multimodale peut également s'y ajouter, précisant ainsi la transcription avec des parties de la vidéo. Dans une perspective interactionniste, nous devons nous concentrer sur toute la séquentialité, soit le contexte d'apparition dans le tour de parole (position initiale-médiane-finale) mais aussi dans la séquence interactive (initiation d'un topic-clôture-requête).

Résultats

La littérature autour de la locution *en vrai* est encore maigre. Une recherche sur un corpus écrit (Frantext), décrit l'évolution d'un emploi de "vrai" comme marqueur discursif durant la deuxième moitié du XIXe (Lefeuvre, 2021). L'auteure conclut que l'emploi de *vrai* est absent des corpus oraux spontanés de nos jours mais qu'il pourrait se retrouver dans d'autres segments tels que *c'est vrai que, en vrai.* Au stade actuel de la recherche nous avons quantifié les

1 Ici nombre d'heures exploitées du corpus comme il est toujours en cours.

occurrences disponibles de *en vrai.* En constatant l'essor et l'installation de ce marqueur assez récemment, il nous semblait pertinent d'en faire un objet d'étude avec notre corpus. Sa mobilisation de plus en plus récurrente dans les pratiques langagières pourrait témoigner d'une grammaticalisation ou pragmaticalisation de la locution. De plus, comme le corpus Pauscaf3 est en phase de finalisation, nous n'avons pas encore accès à toutes les occurrences du corpus. A titre comparatif nous avons regardé les occurrences disponibles dans les corpus CFPP, CFPB et CLAPI. Le corpus CFPP comprend 12 occurrences et CLAPI 4 occurrences. Ces résultats corroborent une apparition récente et en cours de l'utilisation de *en vrai.*

- Les premières analyses montrent tout d'abord une augmentation des occurrences de en vrai de manière longitudinale. De plus, il est intéressant de constater que dans ces occurrences, l'usage littéral de en vrai, lié à la réalité empirique, est très peu utilisé2 dans nos corpus.
- Une des fonctions du marqueur en position initiale seraient l'introduction d'un point de vue subjectif du.de la locuteur.trice. Il peut aussi être utilisé comme marqueur du discours afin de prendre la parole. A l'inverse, une position finale montre une recherche d'affiliation souvent couplée avec le regard. Une analyse multimodale est essentielle à une compréhension plus fine des fonctions. En effet, la prise en compte de la prosodie est également un élément crucial à explorer, à savoir qu'une prosodie montante sur une occurrence en début de tour pourrait projeter un nouveau topic tandis qu'une occurrence en fin de tour avec une prosodie descendante jouera plus un rôle conclusif ou de recherche d'affiliation.
- De manière plus intuitive, *en vrai* marquerait une sorte de contraste entre un point de vue général et son point de vue subjectif, rendant le.la locuteur.trice responsable de son propre énoncé. Il est de ce fait assez intéressant que nous l'ayons trouvé dans des contextes d'auto-correction afin de rectifier son propre discours. Nous repérons également des similitudes d'usage avec *en fait*. Serait-il possible que les deux locutions co-existent jusqu'à ce qu'une finisse par supplémenter l'autre ?

Conclusion

Notre recherche sur le marqueur *en vrai* est encore aux prémices de son analyse mais les résultats actuels laissent présager une augmentation du marqueur dans les pratiques langagières. Le fait d'adopter une perspective longitudinale avec des données récentes nous a permis de saisir certains des usages actuels et de remarquer sa progression. Avec les outils de l'analyse conversationnelle, nous avons pour objectif de déceler les différentes fonctions accomplies par le marqueur et peut-être repérer à quel stade de son évolution il se trouve.

2 Dans le corpus CLAPI nous avons trouvé ce type d'occurrence : « toi t'es nul au foot en vrai mais pas à ça » le « pas à ça » référant à jouer au foot via les jeux vidéo.

Références bibliographiques

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G.H. Lerner (Ed.), Conversation Analysis. Studies from the first generation. John Benjamins, p. 13-31.

Mondada, L. (2017). Nouveaux défis pour l'analyse conversationnelle : l'organisation située et systématique de l'interaction sociale. *Langage et société*, *160161*(2), 181-197.

Mondada, L. (2008). Contributions de la linguistique interactionnelle. In *Congrès mondial de linguistique française* (p. 073). EDP Sciences.

Lefeuvre, F. (2021). Vrai comme marqueur discursif. Marques d'oralité et représentation de l'oralité en français. (halshs-03143458).

Pekarek, S. (2005). Grammaire–Discours–Interaction: vers une approche interactionniste des ressources grammaticales liées à l'organisation discursive. *Travaux neuchâtelois de linguistique*, (41), 1-14.

Psathas, G. (1994). *Conversation analysis: The study of talk-in-interaction*. Sage publications.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turntaking for conversation. Language, 50(4), 696-735.

La fouille de motifs comme outil de détection des genres de discours

Hugo Dumoulin[1] & Timothée Premat[1,2]
[1] Modyco, Université Paris Nanterre
[2] Ceditec, Université Paris-Est Créteil

Introduction

Il est souvent proposé que les routines discursives soient des observables caractéristiques des genres de discours (Née et al. 2016, Sitri & Véniard 2017). À l'appui de cette thèse, on avance régulièrement, et à bon droit, la définition bakhtinienne des genres de discours comme des "types relativement stables d'énoncés" (1984). Pourtant, peu de travaux s'attachent à démontrer la thèse du point de vue empirique. Nous proposons de fournir ici un argument formel à l'appui de la caractérisation bakhtinienne des genres de discours comme "types relativement stables d'énoncés".

Nous faisons le choix de modéliser les unités phraséologiques du discours comme des motifs séquentiels (Srikant & Agrawal 1996) (dorénavant MS). Dans ce cadre, nous proposons d'utiliser la fouille de motifs comme outil de détection des genres de discours. Le protocole expérimental est le suivant : à partir d'un corpus annoté morpho-syntaxiquement issu de quatre genres de discours relativement proches, nous entraînons des classifieurs du type machine à vecteur de support (SVM) à prédire le genre de discours de chaque texte. Nous comparons la performance du classifieur auquel nous donnons une représentation du corpus sous la forme de motifs séquentiels, à celle de deux classifieurs témoins pour qui le corpus est représenté sous formes de lemmes, ou de pos. Nous observons que le classifieur qui a accès à des données sous forme de motifs séquentiels dispose d'une bonne performance à prédire le genre du discours ; les MS permettent une meilleure classification des textes que les part-of-speech (pos), et une classification presque aussi bonne que celle fondée sur les lemmes. Les MS échappant à la dimension référentielle des textes, qui peut être contingente à la définition linguistique de leur généricité, cette recherche apporte un argument formel à l'appui de la thèse bakhtinienne.

Corpus et méthodologie

Corpus

Notre corpus est composé de quatre sous-corpus issus de différents travaux actuels en analyse du discours[1] : (a) 38 rapports d'auto-évaluation d'unités de recherche de l'Université Paris Nanterre en 2018 (Lethier et al. 2022), (b) 368 comptes-rendus de Conseil d'Administration (CR_CA) de la même université de 1984 à 2018 (Diwersy et al. 2024), (c) 141 articles scientifiques issus du corpus *Scientext* (Tutin & Grossman 2014) et (d) 363 comptes-rendus de la session 2018 de l'Assemblée Nationale issus du corpus *Parlamint* (CR_AN ; Erjavec *et al.* 2022). Ces textes contrastent en genre (rapport, CR, articles), en sphère d'activité (académique, politique), et en date (récents sauf (b), longitudinal sur 34 ans). Ils totalisent 24 365 285 mots, dont la ventilation est donnée en table 1 ; les mesures et modèles utilisés ne sont pas sensibles

1 Nous remercions les auteurs de ces travaux pour les avoir mis à notre disposition.

à la différence de taille des sous-corpus. Tous les textes sont de langue française et ont été annotés avec Stanza (Qi et al 2020) et CamemBERT (Martin 2020) ; toute annotation préexistante a été ignorée.

Sous-corpus	Genre	Sphère d'activité	Nb. textes	Nb. mots
a) Rapports	Rapport	académique	38	917 593
b) CR_CA	Compte-rendu	académique	368	2 551 571
c) Scientext	Article_scientifique	académique	141	9 262 055
d) CR_AN	Compte-rendu	politique	363	11 634 066

table 1. : Composition du corpus

Méthodologie

S'agissant de l'extraction de MS, notre chaîne de traitement repose sur une adaptation du travail de Mekki (2022) : les extracteurs PrefixConstraint (Pei et al. 2001) et BIDESpanTree (Wang & Han 2004) sont utilisés pour extraire les MS dont la fréquence relative dépasse un certain seuil minsup (pourcentage de phrases du corpus contenant le motif), et correspond à une contrainte de nombre d'itemsets minimal fixée. Lors de l'extraction des MS, chaque mot est représenté comme un itemset multidimensionnel (Mellet & Longrée 2012) : c'est-à-dire à la fois une forme, un lemme, une étiquette morphosyntaxique, une relation de dépendance, *etc*. Par exemple, pour l'extraction de MS, la phrase *Le président arrive* est représenté par la séquence de 3 itemsets :

{lemma=le, POS=DET, dep=det} {lemma=président, pos=NOUN, dep=nsubj} {lemma=arriver, pos=V, dep=root}

Ainsi les MS extraits seront des unités phraséologiques qui combinent dans une même séquence syntagmatique différents niveaux de description linguistique (Mellet & Longrée, 2012 ; Longrée et Vanni 2025). Avec CloSPEC (Béchet et al 2015), on réduit la masse des données en filtrant les motifs fréquents obtenus pour ne retenir que les motifs clos[2].

Une fois les motifs extraits, on utilise le langage de requêtes CQP au sein de l'environnement de l'IMS Corpus Work Bench (Evert & Hardie 2011) pour obtenir la fréquence de chaque motif dans chaque texte de notre corpus. Ainsi, pour un seuil de minsup fixé, l'on obtient un tableau de contingence qui ventile la fréquence des motifs par texte, lequel peut être interprété à l'aide d'analyses factorielles ou de calcul de spécificités. Ainsi, à la différence de plusieurs travaux actuels, l'on ne réalise pas de pré-traitement de motifs comme la neutralisation des formes pleines (Legallois, Charnois & Poibeau, 2016), ni de post-traitement interprétatif (Longrée et Mellet 2018).

En revanche, nous ajoutons un post-traitement qui relève de l'apprentissage supervisé, en entraînant une série de classifieurs SVM à prédire les genres discursifs des textes du corpus sur la base du tableau de contingence motifs/texte précédemment obtenu. On sélectionne le meilleur classifieur à travers une méthode de gridsearch qui optimise exhaustivement les hyperparamètres du modèle (paramètre C, noyau linéaire ou radial, paramètre gamma) et on évalue les résultats de ce classifieur à travers une k-fold cross-validation (5 splits, 10 répétitions). On compare les performances avec celles qui sont obtenues en partant d'un tableau de contingence lemmes/texte et pos/texte, en retenant au maximum autant de types de lemmes différents (ou pos) que de types de motifs différents (c'est-à-dire autant de colonnes que dans le tableau de contingence des motifs).

2 Motifs qui ne sont contenus dans aucun motif plus large qui serait contenu dans un plus grand nombre de phrases

Résultats

Pour un classifieur donné, on représente la frontière de décision entre les quatre classes prédites par l'intermédiaire d'une analyse des composantes principales (ACP) :

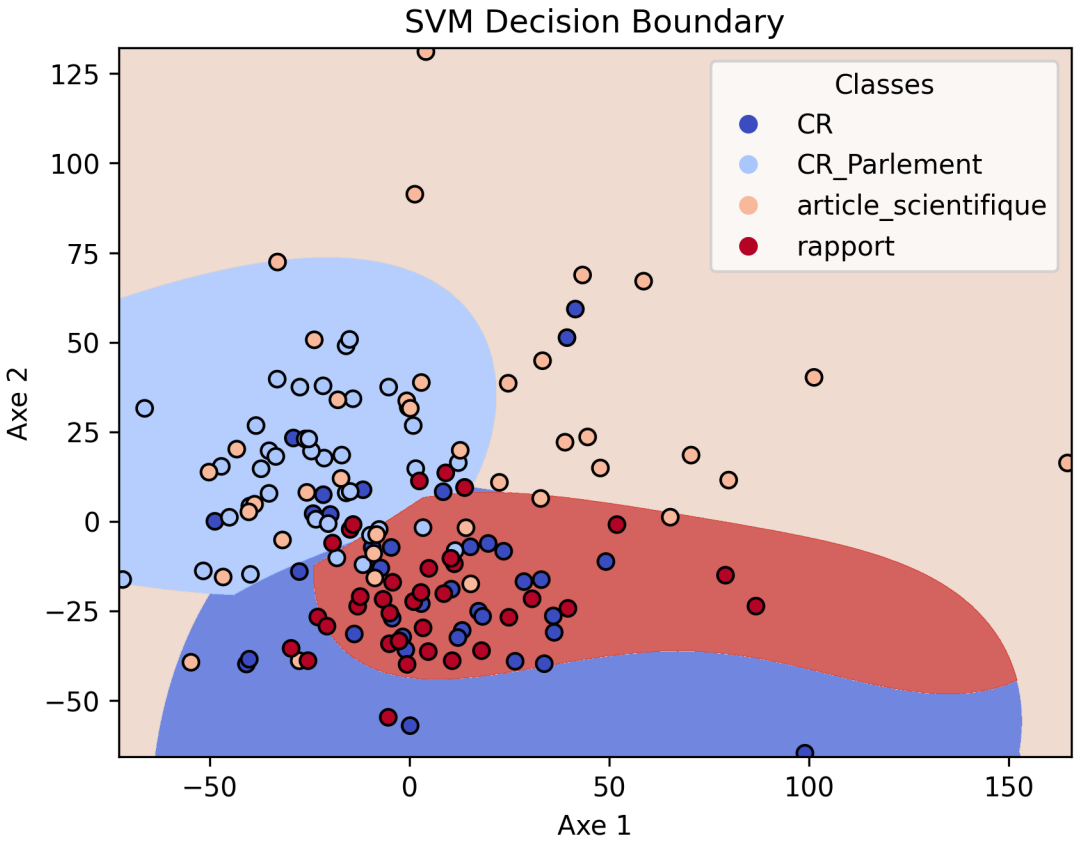


figure . 1 ACP et frontière de décision pour le classifieur radial (C=10, gamma=0,0001) des textes en fonction de la fréquence des motifs (minsup 5)

On compare la f-mesure globale (pondérée par le poids respectif de chaque classe) du meilleur classifieur pour chaque type de descripteurs (motifs, lemmes, pos). On observe que les classifieurs à motifs sont globalement meilleurs que les pos et presque aussi bons que les lemmes dans la tâche de prédiction des genres. L'on ne trouve pas de motifs à 5 itemsets minimum pour un minsup 25.

	motifs	100lemmes	15pos
minsup 25, itemset min = 3	0,93	0,98	0,94
minsup 10, itemset min = 3	0,95	0,98	0,94
minsup 5, itemset min = 3	0,95	0,98	0,94
minsup 25, itemset min = 5			
minsup 10, itemset min = 5	0,82	0,98	0,94
minsup 5, itemset min = 5	0,90	0,98	0,94

table 2. : f-scores globaux pondérés du meilleur classifieur en fonction de différents paramétrages de la fouille de motif (minsup, itemset_min)

Ce bon score des classifieurs qui prennent pour base les motifs appuie l'argument de leur pouvoir discriminant pour une tâche de détection des genres. De ce fait, il y a lieu de parler ici d'un argument formel et empirique à l'appui de la définition bakhtinienne des genres comme

“types relativement stables d’énoncés”, dans un cadre où l’on a modélisé les unités phraséologiques comme des motifs séquentiels. Les lemmes restent cependant porteurs d’une information référentielle qui semble pour l’instant plus discriminante dans l’état de notre modèle.

Références bibliographiques

Bakhtine, M. (1984). *Esthétique de la création verbale*. Gallimard.

Béchet, N., Cellier, P., Charnois, T. & Crémilleux, B. (2015). Sequence mining under multiple constraints. In: *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, 908-914.

Longrée, D. & Vanni, L. « Identification des motifs textuels. Entre statistique et *deep learning* », *Corpus* [En ligne], 27 | 2025, mis en ligne le 13 mai 2025, consulté le 16 mai 2025. URL : http://journals.openedition.org/corpus/10326 ; DOI : https://doi.org/10.4000/13woj

Émilie Née, Frédérique Sitri et Marie Veniard, « Les routines, une catégorie pour l’analyse de discours : le cas des rapports éducatifs », Lidil [En ligne], 53 | 2016, mis en ligne le 01 janvier 2017, consulté le 28 février 2024.

Erjavec, T. *et al.* (2022). The ParlaMint corpora of parliamentary proceedings. *Lang. Resour. Eval.* 57, 1 (Mar 2023), 415–448. https://doi.org/10.1007/s10579-021-09574-0

Evert, Stefan and Hardie, Andrew (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

Legallois, D, Charnois, T. et Poibeau, T. (2016). Repérer les clichés dans les romans sentimentaux grâce à la méthode des « motifs ». *LIDIL - Revue de linguistique et de didactique des langues*.

Legallois, D. (2006), “Des phrases entre elles à l’unité réticulaire de textes”, *Langages* 163, 56-70. DOI : 10.3917/lang.163.0056

Lethier, V., Née, E., & Dumoulin, H. (2022). Caractériser les logiques disciplinaires et institutionnelles dans les rapports d’autoévaluation. *JADT 2022 : actes des 16èmes Journées internationales de l’Analyse statistique des Données Textuelles*.

Longrée, D. & Mellet, S. (2018). Towards a topological grammar of genres and styles: a way to combine paradigmatic quantitative analysis with a syntagmatic approach. In The Grammar of Genres and Stylesn: From Discrete to Non-Discrete Units, edited by Dominique Legallois, Thierry Charnois, and Meri Larjavaara, 140–163. Berlin/Boston: de Gruyter.

Martin, L. *et al.* (2020). CamemBERT: a Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.645

Mekki, J. (2022). *Caractérisation de registres de langue par extraction de motifs séquentiels émergents*. Thèse de doctorat, Université de Rennes.

Mellet, S. et Longrée, D. (2012). Légitimité d'une unité textométrique : le motif. In A. Dister, D. Longrée, G. Purnelle (éds.), *Actes des Journée d'analyse des données textuelles 2012*, 715-728.

Qi, P. *et al*. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Association for Computational Linguistics (ACL) System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Sascha Diwersy, Hugo Dumoulin, Caroline Facq-Mellet, Cyrielle Montrichard, Frédérique Sitri (2024). La fac et son temps : Explorations textométriques d'un corpus diachronique de comptes rendus universitaires. *JADT*, 2024, Bruxelles, Belgium.

Sitri, F. et Veniard, M. (2017). Routines discursives, variation et normes de genre. Langage et société, 159(1), 99-114. https://doi.org/10.3917/ls.159.0099.

Srikant, R., Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements. In: Apers, P., Bouzeghoub, M., Gardarin, G. (eds) Advances in Database Technology — EDBT '96. EDBT 1996. Lecture Notes in Computer Science, vol 1057. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0014140

Tutin & Grossman (2014) *L'écrit scientifique : du texte au discours. Autour de Scientext.* Presses universitaires de Rennes.

La production orale comme levier pour la production écrite ? le cas de l'étudiant dyslexique

Matthieu Quignard & Audrey Mazur
UMR 5191 ICAR (CNRS, Université de Lyon, ENS de Lyon, Université Lumière Lyon 2)
matthieu.quignard@ens-lyon.fr, Audrey.Mazur@ens-lyon.fr

Dyslexie et production textuelle

À l'Université, en France, un des troubles les plus représentés parmi les étudiants déclarés à la mission Handicap est la dyslexie-dysorthographie (Beuchon et Bouhours, 2023). Ce trouble fait partie des troubles neurodéveloppementaux et plus spécifiquement des troubles spécifiques des apprentissages (DSM5, 2013). Il se caractérise par des difficultés dans la reconnaissance exacte et/ou fluente de mots ainsi que par une orthographe des mots et des capacités de décodage limitées (Habib, 2018). Pour les individus typiques, le système de conversion phonème-graphème s'automatise vers 12 ans (Berninger et Swanson, 1994), permettant la mobilisation des ressources cognitives pour des processus de haut niveau tel que la révision. Pour les personnes présentant une dyslexie-dysorthographie, ce système de conversion n'est pas totalement opérationnel, même à l'âge adulte. Ainsi, des difficultés importantes persistent en production écrite, impactant la qualité finale du texte (Fayol et Miret, 2005). Même s'ils se développent, les travaux sur l'adulte dyslexique restent rares (Cavalli et al., 2020), et cela est d'autant plus vrai concernant les manifestations de la dyslexie en production langagière spontanée. Ce présent papier propose des analyses sur des productions écrites et orales d'étudiants présentant une dyslexie-dysorthographie et contrôles. Les individus ont produit à la fois à l'oral et à l'écrit, deux types de textes : un narratif et un expositif. De nombreux indicateurs linguistiques permettent d'aborder ces textes, que ce soit lexicaux, orthographiques, de longueurs ou encore de fréquences. Nous proposons d'aborder ce corpus multivarié selon la perspective de l'ordre de passation pour étudier les potentiels effets d'amorçage (*priming effect*), d'une modalité vers une autre (oral vs. écrit).

Question de recherche et hypothèses

L'ordre de production des textes a été contrôlé dans la mesure où des études ont révélé un effet significatif pour certains critères d'analyse dans les productions de francophones natifs typiques (entre autres, Jisa, 2004). Produire à l'écrit avant une production orale peut impacter positivement cette dernière, les individus activant et mobilisant des formes de l'écrit, transférées ensuite à l'oral, comme des phrases complexes (Gayraud, Jisa et Viguié, 2001) ou des syntagmes nominaux lexicaux (Mazur-Palandre, 2015). Produire à l'oral peut également avoir un effet positif sur le texte écrit qui suit. En effet, des analyses révèlent que les textes des individus ayant produit dans l'ordre oral-écrit ont des unités-terminales (clause principale et ses subordonnées) plus longues que les textes des individus ayant produit dans l'ordre écrit-oral (Mazur-Palandre et Jisa, 2016). Mais quel impact peut avoir l'ordre de production sur les étudiants présentant une dyslexie-dysorthographie ? Nous pouvons faire l'hypothèse que les textes écrits des étudiants présentant une dyslexie-dysorthographie ayant produit dans l'ordre oral-écrit pourraient bénéficier du passage antérieur à l'oral. La modalité orale n'impliquant

pas la contrainte du code orthographique, entre autres, la gestion des ressources cognitives des étudiants présentant une dyslexie-dysorthographie se verrait allégée de cette contrainte, ce qui pourrait être bénéfique, par exemple, en produisant des textes plus longs. Ainsi, nous supposons que *le passage à l'oral avant l'écrit pourrait aider quant à la gestion de certains processus, facilitant alors la mise à l'écrit.*

Méthodologie

Dans le cadre de projets[1], les participants ont été sélectionnés suite à la diffusion d'un questionnaire à l'Université de Lyon. Après un bilan neuropsychologique et orthophonique[2], il a été demandé à une quarantaine d'étudiants de produire quatre textes, à l'écrit et à l'oral. Les textes de 21 étudiants dyslexiques et 22 contrôles appariés en genre, âge et niveau d'étude ont été collectés. Les individus dyslexiques-dysorthographiques, diagnostiqués durant l'enfance, ont suivi au cours de leur enfance/adolescence des rééducations. Tous les individus sont francophones natifs et monolingues, ayant suivi le parcours éducatif en France. Les critères d'exclusion ont écarté tout individu présentant des déficits auditifs, visuels ou d'autres troubles. Les données ont été collectées dans le but d'obtenir des données *off-line* (le texte en lui-même) et des données *on-line*, à savoir chronométriques (pauses, débit, etc.), non abordés dans cette étude. Les individus ont été invités à produire quatre textes, dans un contexte semi-expérimental (thématique imposée et texte élicité à partir d'une courte vidéo muette) : narratif oral, narratif écrit, expositif oral et expositif écrit. La collecte s'est organisée en deux sessions séparées d'une semaine. Lors de la session S1, les individus ont visionné une vidéo d'élicitation puis produit deux textes (narration ou exposition, à l'écrit et à l'oral) séparés par une tâche de distraction. Lors de la session S2, ils ont produit deux autres textes. L'ordre de production a été contrôlé : la moitié produisait un texte écrit puis un oral et l'autre moitié un texte oral puis un écrit.

Corpus

Le corpus ainsi collecté permet alors d'avoir pour chaque individu quatre textes sur une même thématique, permettant alors des comparaisons linguistiques. Ce corpus reste un échantillon de ce que nous pourrions observer par ailleurs mais dimensionné de sorte à pouvoir répondre à des questions de recherche en psycholinguistique. Il est ainsi composé de 172 textes, 86 oraux et 86 écrits. La distribution (moyenne et écart-type) des nombres de mots est donnée dans le tableau 1. L'ensemble du corpus comporte 43 662 mots (la ponctuation ayant été exclue).

	Expo. oraux	**Expo. écrits**	**Narr. oraux**	**Narr. écrits**
Étudiants Dys. (21)	301 (266)	198 (102)	368 (235)	211 (133)
Étudiants Cont. (22)	281 (203)	183 (139)	305 (205)	186 (113)

table 1. : Nombre moyen de mots par texte selon les groupes, genres et modalités

Le protocole expérimental a été pensé afin d'obtenir un corpus comparable, c'est-à-dire divisible en parties appariées : des sujets appariés (dyslexiques / contrôles), des types de textes appariés (expositif / narratif), des modalités de productions appariées (orale / écrite) et un ordre de production apparié (oral puis écrit / écrit puis oral). Le corpus est donc interrogeable selon de multiples dimensions que l'on peut choisir de croiser ou non. Nous nous focalisons sur les effets d'ordre ou effet d'amorçage. Ainsi, concernant l'amorçage par la modalité orale, nous

1 Projets PEPS-CNRS ETUDYS et DYS'R'ABLE, co-financés par le LabEx ASLAN, les laboratoires DDL (UMR 5596 - CNRS et Université Lyon 2) et ICAR (CNRS, Université Lyon 2 et ENS de Lyon).

2 Pour plus d'informations sur ces deux premières étapes du protocole, cf. Mazur et Quignard, 2023

comparons par ANOVA (ou test de Kruskal-Wallis) la moyenne d'un indicateur donné sur les textes écrit (genres confondus) selon que ces textes ont été produits avant ou après les textes oraux. Et de façon analogue, nous comparons la moyenne d'un indicateur donné sur les textes expositifs (modalités orale/écrite confondues) selon ces textes ont été produits avant ou après un texte narratif. Dans le cadre de cette étude exploratoire, nous retiendrons deux indicateurs globaux (*nombre de mots* et de *lemmes différents*) ; deux indicateurs de longueur (*nombre de lettres* et de *phonèmes*) ; deux indicateurs de fréquences (*la fréquence des mots dans un corpus de littérature vs. la fréquence des mots dans les sous-titres de films*) ; et enfin trois indicateurs de complexité : la *complexité articulatoire* (mesure de l'écart à l'alternance stricte de consonnes et de voyelles), la *cohérence phonème-graphème* qui mesure la complexité à écrire certains mots et *la cohérence graphème-phonème* qui mesure la complexité à lire certains mots. Ces indicateurs sont issus des bases lexicales Lexique.org (New et al., 2004) et Manulex (Lété et al., 2004). En plus de ces indicateurs, nous avons également testé l'effet de l'ordre sur le *nombre d'erreurs orthographiques* (lexicales et grammaticales) dans les textes écrits.

Nos hypothèses opérationnelles sont donc les suivantes : les étudiants présentant une dyslexie ayant produit un texte oral avant un texte écrit produisent des textes écrits : contenant plus de mots et de lemmes différents (H1) ; avec des mots plus longs en nombre de lettres et phonèmes (H2) ; des mots moins fréquents (H3) ; et des mots ayant une complexité articulatoire et une opacité orthographique plus élevées (H4) ; présentant moins d'erreurs (H5).

Premiers résultats et perspectives

Nous avons réalisé, pour chacun des indicateurs listés ci-dessus, un test de Kruskal-Wallis (nombre d'étudiants présentant une dyslexie = 21 ; nombre d'étudiants contrôles = 22). Les premiers résultats, sur les indicateurs généraux, ne révèlent aucun effet de l'ordre de production que ce soit chez les étudiants présentant une dyslexie ou chez les étudiants contrôles (nos hypothèses H1 à H4 ne sont donc pas vérifiées). En revanche, pour l'indicateur du nombre d'erreurs, les analyses révèlent un effet significatif seulement sur les étudiants présentant une dyslexie ($H_{(1,20)} = 4.25$; $p = 0.0392$).

En effet, les étudiants présentant une dyslexie, ayant produit un texte oral avant leur texte écrit, ne produisent pas de texte : 1- qui contiennent plus de mots et de lemmes différents (H1) ; 2- qui sont plus longs (en nombre de lettres et phonèmes – H2) ; 3- qui contiennent des mots moins fréquents (H3) et 4- ou qui présentent une complexité articulatoire et une opacité orthographique plus élevées. Cela est observable également pour le groupe contrôle. Ainsi, les différences concernant le nombre de mots et de lemmes différents, la longueur, la fréquence, la complexité articulatoire et l'opacité orthographique ne sont pas significatives, selon que les textes écrits aient été produits avant ou après la production d'un texte oral.

Toutefois, quand nous nous intéressons à des indicateurs plus fins, tels que le nombre d'erreurs, nous pouvons observer un effet significatif de l'ordre de production. Les étudiants présentant une dyslexie, ayant produit un texte oral avant leur texte écrit, réalisent beaucoup moins d'erreurs dans leur texte écrit que les étudiants présentant une dyslexie qui n'ont pas réalisé de texte oral avant leur écrit (10,7 erreurs par texte contre 17,9). Ce résultat n'est pas observé chez les étudiants contrôles : cette sensibilité à l'ordre de production est donc une spécificité des étudiants présentant une dyslexie. Ces résultats suggèrent alors que le fait d'activer des formes linguistiques à l'oral faciliterait leur mise à l'écrit, au moins en ce qui concerne l'orthographe.

Ce résultat encourage à aller tester d'autres indicateurs plus fins tels que les révisions, les pauses ou encore des indicateurs de vitesses d'écriture (débit). Des études plus approfondies du lexique

et d'indicateurs *on-line* et *off-line* permettraient de savoir s'il pourrait être pertinent d'envisager des activités orales avant toute production écrite en remédiation ou en contexte scolaire.

Références bibliographiques

Afonso, O., Suárez-Coalla, P., and Cuetos, F. (2015). Spelling impairments in Spanish dyslexic adults. *Frontiers in Psychology*, 6, 466.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders DSM-5 (5e éd.). Arlington, VA: American Psychiatric Publishing.

Berninger, V., and Swanson, H. (1994). Modifying Hayes and Flower's model of skilled writing to explain beginning and developing writing. In J. Carlson, and E. Buttterfly (Ed.). *Advances in Cognition and Educational Practice, Children's Writing: Toward a Process Theory of the Development of Skilled Writing*, Vol. 2 (pp.57-8). Greenwich: J.A.I. Press.

Beuchon, T. et Bouhours, A. (2023). Les étudiants en situation de handicap dans l'enseignement supérieur. In P. Schuhl (Ed.) *L'état de l'Enseignement Supérieur, de la recherche et de l'Innovation en France* (pp. 38-39). Paris : Sous-direction des systèmes d'information et des études statistiques Ministère de l'Enseignement supérieur et de la Recherche (SIES).

Cavalli, E., Duncan, L. et Colé, P. (2020). La dyslexie chez l'adulte : une introduction. In P. Colé, E. Cavalli et L. Duncan (Eds.), *La dyslexie à l'âge adulte, Approche neuropsychologique*, 1-18. Louvain-la-Neuve : De Boeck Supérieur.

Connelly, V., Campbell, S., MacLean, M. et Barnes, J. (2006). Contribution of Lower Order Skills to the Written Composition of College Students With and Without Dyslexia. *Developmental neuropsychology*, 29, 175-196.

Farmer, M., Riddick, B., and Sterling, C. (2002). *Dyslexia and inclusion: assessment and support in higher education*. London and Philadelphia: Whurr Publishers.

Fayol, M., and Miret, A. (2005). Écrire, orthographier et rédiger des textes. Psychologie *Française, 50*, 391-402.

Gayraud, F., Jisa, H. et Viguié, A. (2001). The development of syntactic packaging in French children's written and spoken texts. *Developing Literacy Across Genres, Modalities, And Languages* 1:169-181

Gregg, N., Coleman, C., Davis, M. et Chalk, J. C. (2007). Timed essay writing: Implications for high-stakes tests. *Journal of Learning Disabilities, 40*(4), 306–318.

Jisa, Harriet (2004). Growing into academic French, Later Language Development: Typological and psycholinguistic Perspectives. *TILAR, 3, 135-162.*

Habib, M. (2018). Dyslexie de développement. *EMC, Psychiatrie/Pédopsychiatrie, 0(0),* 1-12.

Hatcher, J., Snowling, M., and Griffiths, Y. (2002). Cognitive assessment of dyslexic students in higher education. *British Journal of Educational Psychology, 72*, 119–133.

Leloup, G., Launay, L. and Witko, A. (2022). Argumentaire scientifique et clinique, In Recommandations de Bonne Pratique d'Évaluation, de Prévention et de Remédiation des troubles du langage écrit chez l'enfant et l'adulte. Méthode : Recommandations par consensus

Formalisé, Collège Français d'Orthophonie, pp.7-190, consultation : https://www.college-francais-orthophonie.fr/wp-content/uploads/2022/03/RECOS_LE.pdf

Lété, B., Sprenger-Charolles, L., Colé, P. (2004). Manulex: A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments, & Computers*, 36, 156-166.

Mazur-Palandre, A. (2015). Overcoming Preferred Argument Structure in written French: development, modality, text type. *Written Language and Literacy*. 18(1), 25-55.

Mazur-Palandre, A. et Jisa, H. (2016). Maintenir l'information sous forme lexicale : un indice de maturité linguistique ?, *Congrès Mondial de Linguistique Française - CMLF 2016*. Tours: Institut de Linguistique Française, Psycholinguistique et acquisition.

Mazur, A. et Quignard, M. (2023). French Students with Dyslexia Facing the Punctuation System: Insecurity, Inventory Use, and Error Studies. *Brain Sciences*, 13 (4), pp.532.

New, B., Pallier, C., Brysbaert, M., Ferrand, L. (2004). Lexique 2 : A New French Lexical Database. *Behavior Research Methods, Instruments, & Computers*, 36 (3), 516-524.

Piolat. A. (2004). Approche cognitive de l'activité rédactionnelle et de son acquisition. Le rôle de la mémoire de travail. *LINX Linguist. Inst*., 51, 55–74.

Singleton, C.H. (2001). *Dyslexia in Higher Education: Policy, Provision and Practice. Report of the National Working Party on dyslexia in higher Education*. Hull: University of Hull.

Sterling, C., Farmer, M. Riddick, B., Morgan, S., & Matthews, C. (1998). Adult dyslexic writing. *Dyslexia, 4,* 1-15.

Sumner, E. and Connelly, V. (2020). Writing and Revision Strategies of Students With and Without Dyslexia. Special Series: The Interaction of Reading, Spelling and Handwriting Difficulties with Writing Development–Part 2, 189-198.

Tops, W., Callens, C., van Cauwenberghe, E., Adriaens, J., & Brysbaert, M. (2013). Beyond spelling: The writing skills of stu- dents with dyslexia in higher education. Reading and Writing, 26(5), 705–720.

Le concept de *convivialité* dans une perspective interculturelle en FLE à travers de l'analyse de corpus

Agnieszka Dryjańska[1], Vitalija Kazlauskienė[2]
[1] Université de Varsovie, Varsovie, Pologne
[2] Université de Vilnius, Vilnius, Lituanie

Introduction

La compétence interculturelle, en didactique des langues, implique non seulement la capacité à interagir efficacement avec des locuteurs d'autres cultures, mais aussi à interpréter les réalités linguistiques et discursives à travers le prisme des représentations culturelles qu'elles véhiculent (Byram, 1997, 2021 ; Kramsch, 2013). Elle suppose une prise de conscience des concepts culturels implicites, souvent lexicalisés dans des termes difficiles à traduire directement, comme c'est le cas des mots *convivialité* et *convivial* en français. Cette absence d'équivalence n'est pas uniquement un obstacle linguistique : elle renvoie aussi à des différences de visions du monde, notamment en ce qui concerne la notion de vivre-ensemble, les relations sociales harmonieuses ou les interactions quotidiennes. Dès lors, *convivialité* et *convivial* soulèvent des enjeux lexicaux, culturels et didactiques, en particulier dans les parcours d'enseignement/apprentissage du français langue étrangère (FLE) dans le cadre philologique, où ces lexèmes sont présents dans les manuels, mais insuffisamment exploités de manière conceptuelle. Leur apprentissage en FLE pose un double défi : d'une part, linguistique, puisqu'il s'agit de comprendre des mots à forte charge culturelle ; d'autre part, pragmatique et culturel, car il faut accéder à toutes les pratiques sociales et symboliques qu'ils désignent.

Ces réflexions font suite à nos travaux antérieurs (Kazlauskienė & Dryjańska, 2022 ; 2023), dans lesquels nous avons attiré attention sur l'inadéquation des unités lexicales, notamment des collocations, qui se traduit par la non-coïncidence collocationnelle dans l'expression du concept de *fête* en français, polonais et lituanien. En analysant les lexèmes tels que *jour, journée, fête* dans leur emploi courant et institutionnel, nous avons montré que les équivalents apparents entre langues dissimulent des divergences structurelles et culturelles dans la façon de désigner les événements collectifs ou symboliques. Selon nous, le concept de *convivialité* prolonge et approfondit cette problématique, en tant que mot-concept à haute valeur symbolique, ancré dans la culture interactionnelle propre au monde francophone, et donc très pertinent à analyser dans une perspective de *data driven learning (DDL)*.

Objectifs et méthodologie

Si les mots *convivialité* et *convivial* figurent dans les manuels de FLE, ils restent cependant largement sous-exploités sur le plan lexical et conceptuel et rarement didactisés en tant que véritable notion interculturelle. Or, l'absence d'équivalents directs dans certaines langues comme le polonais et le lituanien - où l'on recourt plutôt à des périphrases approximatives ou à des expressions contextuelles telles que *draugiška atmosfera, towarzyskość* ou *atmosfera życzliwości* – en fait un objet privilégié pour aborder les problématiques interculturelles en FLE.

Dans cette perspective, l'objectif central de cette étude est d'analyser comment enseigner des concepts profondément enracinés dans une culture spécifique, tel que celui de *convivialité*, dont le sens, en tant que construction discursive, sociale et interactive complexe, n'est pas véhiculé dans une autre langue par un simple équivalent lexical.

Notre démarche principale est fondée sur DDL, approche favorisant un travail conceptuel permettant aux apprenants d'appréhender plus efficacement et plus profondément des concepts complexes comme *convivialité*. Nous proposons de recourir à une analyse de corpus qui, s'appuyant sur de nombreux exemples d'usage et des patrons lexicaux, permettra, à travers des analyses quantitative et qualitative, d'esquisser le profil lexico-discursif du mot *convivialité,* corrélat du concept désigné par ce mot. Cette partie de l'étude est basée notamment sur les corpus français suivants : frTenTen23 (*SketchEngine), Leipzig Corpora Collection French et Frantext,* mais également sur des corpus polonais et lituaniens.

Cette partie centrale de notre recherche est précédée par l'examen des quelques méthodes de français, fréquemment utilisées dans les milieux académiques polonais et lituanien, afin de montrer comment les mots *convivialité* et *convivial* y sont présentés. Vu l'insuffisance de leur didactisation, observée dans un manuel, nous explorons les modalités de traduction, la traduction automatique incluse, des mots en question afin d'étudier l'utilité didactique de ces traductions. Le choix de la traduction automatique s'explique par la volonté d'utiliser les mêmes outils auxquels recourent les apprenants, face à des difficultés lexicales.

Notre étude débute par une analyse approfondie du rôle de la convivialité dans la société française, tel qu'il est abordé dans la littérature spécialisée en sociologie, philosophie ou même en politique.

Concept de convivialité

Le processus de conceptualisation correspond à l'activité mentale par laquelle les individus interprètent les informations perçues et organisent leur expérience du monde (Kazlauskienė & Ibrahim, 2023). Il s'agit d'un mécanisme cognitif fondamental par lequel des entités abstraites ou concrètes acquièrent un statut symbolique stable dans la conscience, sous forme de concepts – unités de sens intégrées dans la mémoire culturelle et linguistique (Kubryakova, 2009 ; Evans, 2019).

Dans cette perspective, le concept de *convivialité* semble profondément ancré dans la culture française, à la fois lexicalisé, porteur d'une représentation particulière du lien social et même idéologisé. D'après Priscille de Poncins (2011), la convivialité joue un rôle central dans la pensée sociale contemporaine française, en tant qu'alternative face aux inégalités croissantes et à la crise écologique. Elle incarne une idée de société fondée sur la solidarité et la coopération. Cette approche prolonge la réflexion initiée par Ivan Illich (1973), qui conçoit la convivialité comme une condition essentielle à une société autonome, où les outils, les institutions et les relations humaines ne sont pas aliénants mais orientés vers l'émancipation. Dans une perspective plus critique, Quessada (2003), traitant la convivialité de « fable contemporaine », souligne la polyvalence du terme, devenu omniprésent dans les discours contemporains pour qualifier un idéal consensuel du vivre-ensemble. La convivialité, pensée comme une relation fondée sur le partage et la mise en commun, est un phénomène « entre » (bidem). Quessada rappelle ses origines festives, issues des cercles d'amis, pour insister sur la transformation du mot en horizon politique et éthique, dans la lignée d'Illich. Enfin, les réflexions contemporaines autour du convivialisme cherchent à institutionnaliser cette vision relationnelle et éthique du lien social en l'opposant à l'individualisme néolibéral, et en appelant à la restauration des

relations humaines fondées sur les principes de modération, de reconnaissance mutuelle et de coopération (Caillé, 2011).

Résultats de l'analyse lexicographique et de corpus

Dans le manuel Edito B1, les mots *convivialité* et *convivial* n'apparaissent que 5 fois et dans des phrases ne permettant pas d'en saisir sa portée culturelle. Voici un exemple :

« Tu en apprendras beaucoup sur l'ambiance de travail, la convivialité, les événements. » (piste 16)

Ensuite, étant donné que les dictionnaires *TLFi*, *Le Robert* en ligne et *le Grand Dictionnaire Français - Polonais* ne fournissent pas d'explications satisfaisantes, nous avons procédé à la traduction automatique des exemples proposés par Le Robert en ligne. La traduction dominante fournie par le traducteur *DeepL* est *towarzyskość*. Elle n'est pourtant pas appropriée puisque ce substantif polonais désigne **une qualité d'une personne**. La traduction proposée par l'assistant *Copilot – życzliwość –* rend mieux le sens de *convivialité*. Ce nom désigne cependant également **une qualité humaine**. Or, dans les contextes analysés, *convivialité* ne qualifie pas une personne, mais est une caractéristique d'une situation, d'un moment ou d'un espace « entre » des personnes, comme le montrent les expressions *zones de convivialité*, *moment de convivialité*.

Il est à noter que le nom *życzliwość* est plus fréquent, avec 1 258 occurrences dans le Corpus National de la Langue Polonaise (NKJP) contre 38 occurrences pour *towarzyskość*, ce qui le rapproche du substantif *convivialité* en français.

En lituanien *convivialité* se traduit souvent par *bendravimas* (nom), ce qui désigne de manière générale l'acte de communiquer, d'interagir ou d'entretenir une relation sociale avec autrui. Le mot est sémantiquement neutre et large, couvrant à la fois la communication verbale et non verbale, dans divers contextes : quotidien, professionnel, éducatif, familial, etc.

L'analyse de corpus préliminaire révèle que le sens prototypique des mots *convivialité* et *convivial* est lié au contexte festif et gastronomique. Ce lien est attesté par les cooccurrences coordonnées de l'adjectif *convivial*, telles que : *festif, familial, ludique*, ainsi que par ses collocatifs nominaux *ambiance, repas, soirée*.

Dans le corpus frTenTen23, les collocations de *convivialité* de type NA telles que *convivialité intergénérationnelle, islamo-chrétienne, fraternelle, villageoise, chaleureuse* soulignent la dimension sociale du concept. De même, l'exemple extrait de Frantext : *convivialité entre générations* confirme que la convivialité est un phénomène collectif, perceptible dans des groupes sociaux plus larges.

Par ailleurs, les mots *chaleur* et *chaleureux* figurent parmi les associations lexicales les plus fréquentes des mots en question, comme l'illustrent les exemples suivants : *des cercles chaleureux et conviviaux* (*frTenTen23*) *L'ambiance y est chaleureuse et conviviale* (*Hofbräuhaus, collected on 22/10/2007*). Cela met en avant leur connotation très positive.

Enfin, l'analyse du corpus frTenTen23 révèle l'emploi du substantif *convivialité* dans un contexte informatique représenté par la collocation *interface conviviale* (*frTentTEn23*).

Conclusion

La traduction des mots *convivialité* et *convivial* effectuée dans le cadre de la recherche met en évidence que les équivalents polonais et lituaniens renvoient à d'autres concepts qui ne correspondent pas directement au concept de *convivialité*. De plus, l'analyse d'un manuel révèle que ces divergences ne sont pas systématiquement prises en compte dans les pratiques pédagogiques. Cela confirme la nécessité d'explorer d'autres sources linguistiques, à savoir les corpus de textes, afin de familiariser les apprenants avec des mots-concepts. L'étude de corpus que nous avons menée fournit un matériau linguistique riche, permettant d'immerger l'apprenant dans la complexité et la diversité d'usages linguistico-discursifs des mots *convivialité* et *convivial*. La nécessité d'interpréter ces données peut conduire les apprenants vers une meilleure compréhension du concept de *convivialité*, non pas en tant que qualité d'une personne, mais en tant que phénomène opérant dans un espace social, plus ou moins élargi, « entre » des personnes ou même « entre » des personnes et des outils.

Éléments de la bibliographie

Byram, M. (1997). *Teaching and Assessing Intercultural Communicative*. Multilingual Matters. Clevedon.

Byram, M. (2021). *Teaching and Assessing Intercultural Communicative Competence: Revisited* (2nd ed.). Multilingual Matters & Channel View Publications.

Caillé, A., Humbert, M., Latouche, S., Viveret P. (2011). De la convivialité. Dialogues sur la société conviviale à venir. *Cahiers libres*.

Caillé, A. (2011). *Pour un manifeste du convivialisme*. Le Bord de l'eau, series: « Documents ».

De Poncins, P. (2011). De la convivialité. Dialogues sur la société conviviale à venir La Découverte, *Revue Projet,* 323(4), 99–99.

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.

Dryjańska A., Kazlauskienė V. (2023). Jour, Journée au sens de 'fête' et leurs équivalents collocationnels en polonais et lituanien : défis pour le FLE. *Białostockie Archiwum Językowe*, 23, 101–128.

Evans, V. (2019). *Cognitive Linguistics: A Complete Guide*. Edinburgh University Press.

Illich, I. (1973). *La convivialité*. Seuil.

Kazlauskienė, V., Ibrahim U. (2023). Liberté sans contrainte : extension de l'espace conceptuel du fonctionnement des collocations avec liberté d'aujourd'hui. *Études de Linguistique, Littérature et Art, 65 : Entre liberté et contrainte dans la recherche linguistique*, Peter Lang, 25–37.

Kramsch, C. (2013). *The Multilingual Subject*. Oxford University Press (version Kindle).

Kubryakova, E. S. (2009). About concepts captured by the sign. *Studia Linguistica*, 18, 69–75.

Quessada, D. (2003). La convivialité : une relation sans Autre. Quaderni, 53, 67–80.

Morphème interro-exclamatif ti en français hexagonal et en français québécois (1850-1950)
Tu viens ti ?, c'est ti beau !, je sais ti

Gaétane Dostie[1], Florence Lefeuvre[2]
[1] CRIFUQ, Université de Sherbrooke
[2] CLESTHIA, Université Sorbonne Nouvelle

Problématique, objectifs, corpus d'étude

Le morphème interro-exclamatif *ti*[1] utilisé à date ancienne dans les variétés européennes du français y était toujours usuel au début du XX^e^ siècle (entre autres, Paris 1906 : 276-280 ; Vendryes 1921 : 201-203 ; Foulet 1921 ; De Boer 1926 ; Frei 1929 : 264 ; Bally 1944[2^e^] : 212). À l'époque actuelle, *ti* fait figure de reliquat dans ces variétés diatopiques de langue. En français québécois, le morphème a connu un sort différent. *Ti*, qui était toujours en usage dans les années 1940-1950, a vraisemblablement été rapproché de manière inconsciente du pronom d'adresse *tu* par un processus de réanalyse (Picard 1992 : 69 ; Léard 1995 : 221 ; Léard 1996 : 113). De ce fait, il existe de nos jours dans cette variété diatopique de langue deux formes homonymes : le morphème interro-exclamatif *-tu*, employé dans une question directe totale en (3) et dans un énoncé exprimant le haut degré en (4), existe en parallèle au pronom clitique de deuxième personne du singulier *tu*[2].

(1)
Le Père Paturon, d'une voix fière. – j'pourrais-**ti** r'tourner cheuz nous ? Le juge. – parfaitement.
(G. de Maupassant, 1884, *Contes et nouvelles*, t. 1, p. 175 ; Frantext)

(2)
C'es**t-y** vilain, ici !
(J. et E. de Goncourt, 1861, *Sœur Philomène*, p. 146 ; Frantext)

(3)
S : oui c'est ça Gilles tu fais-**tu** encore du skidoo ↑
(CFPQ, 2006, sous-corpus 1, segment 1)

1 Ce morphème est transcrit de plusieurs façons. Outre ti, on rencontre également les formes suivantes : -ti, t-i, -i-, -t-il, -t-y, ty et y (p. ex. dans c'est-y).

2 Nous avons choisi de faire précéder la particule interro-exclamative d'un trait d'union (cf. -tu) pour éviter la confusion avec le pronom tu.

(4)
Voyons. C'est de plus en plus. C'est**-tu** effrayant, c'est effrayant c'est sûr.
(PFC-Québec, 2011 ; FDLQ)

Par ailleurs, *-tu* se rencontre au sein de phrasèmes interactifs (p. ex. *ça se peut-tu*, *on s'entend-tu*, *tu me niaises-tu*, *ça a-tu du bon sens*, *je (le) sais-tu*) à l'instar de ce que l'on observe pour d'autres morphèmes interro-exclamatifs tel *qu'est-ce que* (p. ex. *qu'est-ce que tu veux*, *qu'est-ce que ça peut faire*, *qu'est-ce que je raconte ;* Tutin 2022). De manière schématique, un phrasème « est une expression multilexémique non librement construisible » stockée dans la mémoire des locuteurs (Iordanskaja et Mel'čuk 2017 : 93). Les phrasèmes interactifs s'apparentent à ce qu'on a appelé diversement ailleurs notamment *énoncés liés* (Fónagy 1982 ; Bidaud 2002), *phrases (figées) situationnelles* (Anscombre 2002, 2012 ; Náray-Szabó 2006 ; Klein et Lamiroy 2016), *phrases préfabriquées des interactions* (Tutin 2019, 2022) et, pour certains d'entre eux, *marqueurs discursifs* (Hansen et Visconti éds 2025), notamment *marqueurs discursifs propositionnels* (Andersen 2007). Une recherche systématique dans le Corpus de français parlé au Québec (CFPQ) et, de manière complémentaire, dans plusieurs autres corpus oraux reflétant le français québécois en usage à la fin du XX^{e} siècle et au début du XXI^{e} siècle a permis d'effectuer un premier inventaire d'une quarantaine de phrasèmes interactifs intégrant *-tu* (Dostie 2024).

Partant de ce qui précède, nous nous proposons ici d'examiner les occurrences de *ti* présentes dans des écrits fictionnels parus entre 1850 et 1950[3]. Notre corpus d'étude a été constitué à partir d'une consultation de la base de données FRANTEXT pour le français hexagonal et du Fichier lexical informatisé du Trésor de la langue française au Québec (TLFQ) pour le français québécois. Le nombre total d'occurrences repérées dans l'ensemble des écrits fictionnels contenus dans ces deux bases de données est de 408 (on en trouve 188 dans les textes français et 220 dans les textes québécois).

Pour orienter notre analyse, nous prendrons appui sur les trois usages typiques de *-tu* relevés précédemment. Nous chercherons donc à caractériser les emplois où *ti* était utilisé, en premier lieu, dans des énoncés servant à poser des questions, en deuxième lieu, dans des énoncés exprimant le haut degré et, enfin, en troisième lieu, dans des phrasèmes interactifs et dans quelques constructions lexico-syntaxiques récurrentes. Chacun de ces emplois présente un intérêt particulier dont nous discutons brièvement ci-dessous. Ils sont particulièrement bien documentés pour *-tu* dans les contextes interrogatifs (entre autres, Elsig et Poplack 2006 ; Elsig 2009 ; Villeneuve 2020 ; Bally 2022 ; Beaulieu 2024 ; Bergeron, Dostie et Lefeuvre 2024) et, dans une moindre mesure, dans les contextes exclamatifs (Dubois 2000 ; Bertrand 2014). La question phraséologique est, quant à elle, relativement nouvelle.

3 Le choix de la périodisation pour la conduite de l'étude repose sur les considérations suivantes. La limite temporelle inférieure (1850) se justifie par des raisons pratiques : les premiers journaux et les premiers textes littéraires québécois datent du début XIXe siècle. La limite temporelle supérieure a été fixée à 1950 du fait que le passage progressif de ti à -tu était déjà entamé en français québécois à cette période (§1).

Usages typiques de *ti* dans les écrits fictionnels (1850-1950)

Ti dans les questions

En ce qui concerne l'utilisation de *ti* dans les questions, nous verrons qu'elle est globalement similaire à celle de *-tu* en synchronie (p. ex. combinatoire naturelle aussi bien avec des formes verbales simples que composées), en dépit de quelques différences ponctuelles. À titre d'exemple, nous remarquons que certaines contraintes propres à *-tu* en raison notamment de son association « inconsciente » avec le pronom d'adresse *tu* (§1) ne semblent pas peser sur *ti*. Ainsi, la construction [*vous* V-*tu ?*] est peu fréquente en synchronie comparativement, par exemple, à [*tu* V-*tu ?*] (sans être impossible ; Dostie 2024) ; en revanche, son pendant [*vous* V *ti ?*] ne semble pas inusité, comme l'illustrent (5) et (6).

(5)
- **Vous** avez-**t-y** appelé le curé ? Il est venu…
(L. Hémond, 1916, *Maria Chapdeleine*, p. 210 ; TLFQ)

(6)
De ce pas, je vas lui dire ma façon de penser, au ratichon ! **Vous** venez-**ti**, vous autres ?
(G. Chevalier, 1934, *Clochemerle*, p. 187 ; Frantext)

Ti dans les énoncés exclamatifs exprimant le haut degré

Marandin (2018) distingue trois sous-types d'exclamatives où chacun présente des propriétés spécifiques : les exclamatives scalaires, les exclamatives de manière et les exclamatives à parangon. À partir d'un examen attentif de la combinatoire de *ti*, nous verrons que ce morphème est en lien avec le premier sous-type d'exclamatives. Ainsi, à l'instar d'autres morphèmes qui appartiennent également aux exclamatives scalaires, tels *combien*, *ce que* et *qu'est-ce que*, le morphème *ti* se combine naturellement avec des expressions quantifiables, comme dans (7) et avec des expressions gradables et intensifiables, comme dans (8). *Tu*, en français québécois contemporain, aurait donc conservé les propriétés inhérentes à son unité source. En effet, ce morphème a lui aussi une affinité avec les exclamatives scalaires, comme en attestent les exemples contemporains (9) et (10). Ces derniers sont similaires à ceux présentés en (7) et (8) : sur le plan de la combinatoire lexicale, il y a présence d'expressions quantifiables, gradables et intensifiables.

(7)
- V'la la planche [pour construire les divisions d'un restaurant] qui arrive chez Boucher. **On va-t-y rire!**
(R. Lemelin, 1944, *Pente douce*, p. 125-126 ; Fichier lexical, TLFQ)

(8)
Pas seulement un arbre devant les églises ! Ceux que leurs parents y sont, c'est bien... mais les autres, ils y viennent pour l'hôpital ... Et c'est gai, leurs hôpitaux !... **C'est-y vilain** , ici !... Je suis sûre que je m'en vas être triste encore quinze jours après, chez nous ...
(J. Goncourt, 1861, *Sœur Philomène*, p. 146 ; Frantext)

(9)
mais (.) on a ri↑ (.) <len<**on a-tu ri**>> de Mathieu heille
(CFPQ, 2009, sous-corpus 15, segment 9)

(10)
il **est-tu laid**/ il **est-tu laid**/
(CFPQ, 2011, sous-corpus 26, segment 7)

Ti dans les phrasèmes et dans quelques constructions lexico-syntaxiques récurrentes

En prenant appui sur la liste des phrasèmes interactifs en *-tu* dont nous disposons (§1), nous montrerons que quelques phrasèmes, usuels de nos jours, représentent des variantes formelles « remises au goût du jour » de phrasèmes interactifs existant avec le morphème *ti* à date ancienne. C'est le cas de *je sais-tu* utilisé par le locuteur pour exprimer son ignorance en (11) ('je ne sais pas', 'je n'en ai pas la moindre idée'). Ce phrasème constitue la version « moderne » du phrasème ancien *je sais ti*, attesté au XIX[e] dans l'exemple (12).

(11)
C : pourquoi vous me regardez /
J : ben **je sais-tu** moi (*en hochant la tête négativement*)
(CFPQ, sous-corpus 17, segment 7, p. 96 ; 2009)

(12)
- qué que tu vas faire ?- j'sais**-ti** ?-va-t'en vé l'curé.
(G. de Maupassant, 1886, *Contes et nouvelles*, t. 1, p. 213 ; Frantext)

Enfin, nous prêterons attention à quelques constructions lexico-syntaxiques récurrentes, notamment à [*c'est ti* X] et [*c'est ti pas* X]. La première construction sous-tend des énoncés tantôt interrogatifs (p. ex. *c'est ti vrai ?*, *c'est ti toi ?*), tantôt exclamatifs (*c'est ti drôle*, *c'est ti vilain*). Par comparaison, la seconde construction montre une attirance nette pour l'exclamation (*c'est ti pas malheureux*, *c'est ti pas honteux*) dans les deux variétés de français, mais de manière encore plus marquée en français québécois. Ces différences invitent à se questionner sur la différence sémantique éventuelle qui pourrait exister entre les constructions [*c'est ti* X] et [*c'est ti pas* X] dans l'exclamative.

Bibliographie

Andersen, H L., 2007, « Marqueurs discursifs propositionnels, *Langue française*, 154, p. 13-28.

Anscombre, J.-C., 2000, « Parole proverbiale et structures métriques », *Langages,* 139, p. 6-26.

Anscombre, J.-C., 2012, « Pour une théorie linguistique du phénomène parémique », in : J.-C. Anscombre *et al.* (éds). *La parole exemplaire. Introduction à une étude linguistique des proverbes,* Paris, Armand Colin, p. 21-39.

Bally, A.-S. (2022), « Les interrogatives totales en français québécois dans l'écrit SMS : à la croisée de l'oral et de l'écrit », *SHS Web of Conferences, Congrès Mondial de Linguistique Française* 138, [https://doi.org/10.1051/shsconf/ 202213812006]

Bally, C., 1944 [2[e] édition], *Linguistique générale et linguistique française*, Berne, Francke.

Beaulieu, A., 2024, *Quand la BDQ questionne le français québécois : analyse des structures interrogatives totales dans les bandes dessinées québécoises après 1978*, mémoire de maîtrise, Université de Sherbrooke, 113 p.

Bertrand, A., 2014, « Exclamatives en *-tu*, *donc* et *assez* en français québécois : types et sous-types, mémoire de maîtrise, Université de Montréal, 124 p.

Bergeron-Maguire, M., G. Dostie et F. Lefeuvre, 2024, « Les procédés (morpho-)syntaxiques de l'interrogation directe totale dans les conversations en français québécois des années 2000 : *tu l'as-tu lu?, l'as-tu lu?, est-ce que tu l'as lu?* », *Langue française*, 221 : 1, p. 21-38.

Bidaud, F., 2002, *Structures figées de la conversation : analyse contrastive français-italien*, Bern, P. Lang.

De Boer, C., 1926, « L'évolution des formes de l'interrogation directe en français », *Romania*, 207, p. 307-327.

Dostie, G., 2024, « Phrasèmes interactifs intégrant le morphème interro-exclamatif *-tu* en français québécois parlé (1970-2020). *On s'entend-tu, ça se peut-tu, tu penses-tu, tu veux-tu ben* », 1ère Rencontre annuelle LLCD (Langues et langage à la croisée des disciplines), Atelier « Cooccurrences et marquage discursif » organisé par M. Dargnat et A. Tutin, Sorbonne Université, 9 au 11 septembre 2024.

Dubois, C., 2000, *La grammaire de l'exclamation : aspects théoriques, français de référence et français québéco*is, mémoire de maîtrise, Université de Sherbrooke, 142 p.

Elsig, M., 2009, *Grammatical variation across Space and Time: the French interrogative system*, Amsterdam, John Benjamins.

Elsig, M. et S. Poplack, 2006, « Transplanted dialects and language change: question formation in Québec », *University of Pennsylvania Working Papers in Linguistics*, 12, p. 77-90.

Fónagy, I., 1982, *Situation et signification*, Philadelphia, J. Benjamins.

Foulet, L., 1921, « Comment ont évolué les formes de l'interrogation », *Romania*, 47 :186-187, p. 243-348.

Frei, H., 1929, *La grammaire des fautes*, Paris, P. Geuthner.

Hansen, M. B. M. et J. Visconti (éds), 2025, *Manual of Discourse Markers in Romance*, Berlin, De Gruyter.

Iordanskaja, L. et I. A. Mel'čuk, 2017, *Le mot français dans le lexique et dans la phrase*, Paris, Hermann.

Klein, J.-R. et B. Lamiroy, 2016, « Le figement : unité et diversité. Collocations, expressions figées, phrases situationnelles, proverbes », *L'Information grammaticale*, 148, p. 15-20.

Léard, J.-M., 1995, *Grammaire québécoise d'aujourd'hui. Comprendre les québécismes*, Montréal, Guérin universitaire.

Léard, Jean-Marcel, 1996, « *-Ti / -tu*, *est-ce que*, *qu'est-ce que*, *ce que*, *hé que*, *don* : des particules de modalisation en français », *Revue québécoise de linguistique*, 24 : 2, p. 107-124.

Marandin, J.-M., 2018, *La phrase exclamative et l'exclamation en français contemporain* [En ligne : hal- 01882115].

Náray-Szabó, M., 2006, « Pragmatique et sémantique des phrases figées situationnelles », *Verbum Analecta Neolatin*a, 8 : 2, p. 473-493.

Paris, G., 1906, *Mélanges linguistiques*, Paris, H. Champion.

Picard, M., 1992, « Aspects synchroniques et diachroniques du *tu* interrogatif en québécois », *Revue québécoise de linguistique*, 21 : 2, p. 65-75.

Tutin, A., 2019, « Phrases préfabriquées des interactions : quelques observations sur le corpus CLAPI », *Cahiers de lexicologie*, 114 : 1, p. 63-91.

Tutin, A., 2022, « *Comment dirais-je ?, Que veux-tu ?, Comment ça va ?* », *Lingvisticae Investigationes*, 45 : 2, p. 172-196.

Vendryes, J., 1921, *Le langage. Introduction à l'histoire de la linguistique*, Paris, La Renaissance du livre.

Villeneuve, A.-J., 2020, « Variation stylistique et accommodation langagière : l'interrogation totale en français québécois soutenu », dans D. Bigot, D. Liakin, R. Papen *et al.* (éds.), *Les français d'ici en perspective*, Québec, Les Presses de l'Université Laval, p. 109-130.

Ressources textuelles et corpus

Corpus de français parlé au Québec (2006-2019), Université de Sherbrooke. [https://applis.flsh.usherbrooke.ca/cfpq/]

BAnQ, (*Bibliothèque et archives nationales du Québec*), Gouvernement du Québec. [https://www.banq.qc.ca]

Canadiana, Réseau canadien de documentation pour la recherche. [https://www.canadiana.ca]

FDLQ (*Fonds de données linguistiques du Québec*), Université de Sherbrooke. [https://fdlq.recherche.usherbrooke.ca]

Frantext intégral, CNRS, Université de Lorraine. [Accès institutionnel]

Gallica, BnF (Bibliothèque nationale de France), Ministère de la Culture (France). [https://gallica.bnf.fr/accueil/fr/html/accueil-fr]

Trésor de la langue française au Québec, Fichier lexical informatisé, Université Laval, [https://www.tlfq.org/fonds/]

Les agrégats de marqueurs de discours. Aspects quantitatifs

Mathilde Dargnat [1], Jacques Jayez [2] et
[1] Laboratoire ATILF, Université de Lorraine [2] Laboratoire LORIA, INRIA, CNRS & Université de Lorraine
jjayez@wanadoo.fr, mathilde.dargnat@univ-lorraine.fr

Introduction

Nous entendrons par *marqueurs de discours* (MD) un sous-ensemble des *opérateurs discursifs*, au sens d'Anscombre, Donaire et Haillet (2013, 2018). Ils constituent des *parenthétiques purs* au sens de Koev (2019), qui, à la différence des modaux (du type *probablement*), ou des adverbes illocutoires (du type *franchement*) ne portent ni sur la valeur de vérité ni sur la force illocutoire d'un énoncé. Les MD se divisent en deux grandes catégories selon qu'ils portent une relation de discours ou manifestent l'inscription interactionnelle, cognitive et/ou émotionnelle d'un locuteur dans le discours (Dargnat 2024, Jayez 2025).

Un certain nombre de travaux récents se sont intéressés aux juxtapositions de MD (voir par exemple Crible & Degand 2021, 2024, Cuenca 2024, Dargnat 2022, Dostie 2004, Haselow 2018, Koops & Lohmann 2015, Lohman & Koops 2016), illustrés en français par *mais bon alors*, *donc du coup*, ou *et pourtant*. L'intuition dominante est que les marqueurs de discours (MD) présentent des affinités de proximité et des préférences dans l'ordre de succession et donc que la distribution de leurs juxtapositions n'est pas purement aléatoire. Les travaux existants suggèrent effectivement que les juxtapositions sont soumises à certaines contraintes syntaxiques et sémantico-pragmatiques, mais la taille et la variété des corpus utilisés est limitée et la méthodologie n'est pas uniforme. Nous proposons une approche plus systématique reposant sur une chaîne de traitement à partir de grands corpus écrits et oraux. Notre but est d'examiner certains aspects qui relèvent d'un traitement quantitatif, et qui doivent contribuer à cerner les modes de dépendance ou d'indépendance entre MD co-présents, sans préjuger d'analyses théoriques, qui doivent intervenir dans un second temps.

Chaîne de traitement

Nous avons sélectionné et « nettoyé » 13 corpus; 7 corpus oraux : CFPP (579340 mots), CLAPI (119412 mots), FRA80 (191169 mots), mpf (993956 mots) et TCOF (906779 mots), qui ont été extraits à partir de Ortolang (https://www.ortolang.fr/), ainsi que ESLO (1 et 2) (3494574 mots), qui a été extrait par algorithme à partir du site (http://eslo.huma-num.fr/) et Declics (44814 mots), qui est un corpus non diffusable d'entretiens médicaux (https://lrl.uca.fr/projet_du_labo/projet-declics/). Les 6 corpus écrits sont : deux corpus du Monde Diplomatique (34351 et 1998868 mots) et du Monde (17 020 381 et 18 142 098 mots), achetés auprès du consortium ELRA (https://www.elra.info/), le corpus wikiconflit (505 981 mots, Ortolang) et un extrait du French Reddit disponible à https: //www.kaggle.com/datasets/breandan/french-reddit-discussion/data (75 538 403 mots), soit 119 570 126 mots au total. D'autres corpus sont en cours d'intégration.

Nous avons constitué une liste de MD en partant du dictionnaire de Charlotte Roze (Roze 2013) et en l'étendant à partir d'échantillons de corpus et de notre intuition. À l'heure actuelle, la liste

contient 751 MD, divisés en *connecteurs* (certains adverbes et conjonctions comme *donc* ou *bien que*) et en *particules énonciatives* (PEN, Dargnat 2024). 132 (≈ 18%) de ces MD sont lexicalement ambigus ; par exemple *bon* peut être un adjectif, un nom ou une PEN, *tu parles* peut être une PEN ou une séquence PRO+V comme dans *tu parles chinois*. Il est nécessaire d'annoter les éléments qui pourraient constituer des MD isolés, car leur comptage intervient dans les mesures présentées dans la section 4.

Des tests préliminaires réalisés avec des étiqueteurs probabilistes ou à base de Grands Modèles de Langue (GML) ont montré que la désambigüisation proposée était souvent insuffisante. Par exemple, les emplois de *bien* comme PEN (INTJ dans l'étiquetage) et comme adverbe de manière sont mal distingués. Les MD complexes (*quand même*, *tout à coup*, etc.) sont souvent décomposés. Enfin, les GML souffrent du problème général de la non-explicabilité (effet boîte noire). Nous nous sommes tournés vers une solution à base d'automates finis, lesquels sont lourds à déployer mais transparents et (relativement) faciles à corriger (explicabilité). Celle-ci a été implémentée pour l'essentiel à l'aide d'*Unitex* (https://unitexgramlab.org). L'annotation se fait par transduction en appliquant une séquence (dite *cascade* dans la terminologie d'Unitex) de 611 automates aux corpus qui sont progressivement modifiés par ajout d'annotations comme {parce que,.DMCSUBCONJ} (connecteur conjonction de subordination) ou {tu parles,.DMP} (PEN). Dans la plupart des cas, on élimine les configurations où un élément lexical n'est pas un MD en imposant par exemple un marquage {tu parles,.non-DMP}. L'automate qui s'applique ensuite va vérifier que l'élément n'a pas été marqué ainsi et, selon les cas, le marquer comme DMP ou DMP-CAND, si son statut comme DMP n'est pas certain.

On peut ensuite lancer sur les données (section 3) un certain nombre d'explorations statistiques, décrites en partie dans la section 4.

Données

Les corpus annotés permettent de générer une table contenant toutes les occurrences d'un ou plusieurs MD (juxtaposés) avec actuellement un contexte maximal de 15 mots à gauche et 15 mots à droite. L'ensemble des corpus comprend 7 407 899 (co)occurrences de MD, dont 6 734 173 occurrences isolées (≈ 91%) et 673 726 cooccurrences (≈ 9%).

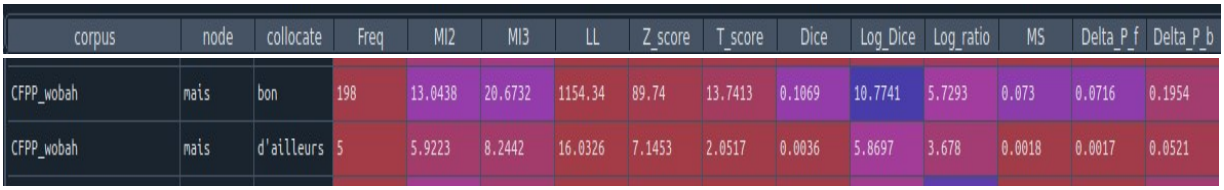
Les cooccurrences peuvent comporter des répétitions (*ah bon ah mais*), au nombre de 72 136, soit 10,7% des cooccurrences, avec une corrélation (modérément : 0,33) positive sur les rangs entre taille de la cooccurrence et taux de répétitions (le quotient de la longueur de la cooccurrence sans répétition par la longueur de la cooccurrence). Pour certains MD, les répétitions peuvent interagir avec les mesures évoquées à la section 4, notamment lorsqu'elles sont contiguës (par exemple *ah ah ah ah*). Les cooccurrences non-contiguës et les éléments notés *euh* (et variantes) dans les transcriptions soulèvent des questions du même genre.

Explorations statistiques

À partir des données de cooccurrences décrites en section 2, plusieurs types d'analyse sont possibles.

La plus classique reste celle des mesures d'association (Brezina 2018, Evert 2005), qui reflètent la « force d'association » entre deux éléments contigus, le pivot (le centre de l'association) et l'associé (son cooccurrent). Brezina (2018) indique que ces mesures s'organisent selon deux dimensions : la fréquence et l'exclusivité. Les mesures sensibles à la fréquence reposent sur la fréquence observée, correspondant au nombre effectif d'occurrences du pivot, de l'associé ou

de combinaisons impliquant un cooccurrent et le pivot, et sur la fréquence attendue, qui se détermine à partir de la présence ou de l'absence du pivot et/ou de l'associé cooccurrent. L'exclusivité estime la tendance pour le pivot et l'associé à figurer davantage ensemble que séparément. Nous avons retenu ici 11 mesures de ce type illustrées dans la figure 1 par la comparaison entre les valeurs pour *bon* et *d'ailleurs* avec *mais* comme pivot, dans le corpus Declics, et un paramètre de longueur maximale des cooccurrences fixé à 10. La colonne *Freq* contient le nombre de juxtapositions (*mais bon*/*mais d'ailleurs*). Les deux dernières colonnes correspondent à deux mesures directionnelles : la ΔP-forward mesure la tendance de l'associé à figurer immédiatement à droite du pivot, et symétriquement pour la ΔP-backward (Schneider 2020).



corpus	node	collocate	Freq	MI2	MI3	LL	Z_score	T_score	Dice	Log_Dice	Log_ratio	MS	Delta_P_f	Delta_P_b
CFPP_wobah	mais	bon	198	13.0438	20.6732	1154.34	89.74	13.7413	0.1069	10.7741	5.7293	0.073	0.0716	0.1954
CFPP_wobah	mais	d'ailleurs	5	5.9223	8.2442	16.0326	7.1453	2.0517	0.0036	5.8697	3.678	0.0018	0.0017	0.0521

figure . 1 Mesures d'association - Une démo

Compte tenu des écarts d'échelle et de nature entre les valeurs fournies par les différentes mesures, une régression utilisant des variables dépendantes numérique (linéaire ou pas) peut susciter certaines réserves. Si l'on adopte le langage de l'expérimentation, les mesures correspondent à des protocoles expérimentaux différents et il n'y a donc pas de raison a priori pour que leurs valeurs soient regroupées, par exemple autour d'une même moyenne. Un modèle linéaire simple ou hiérarchique sur les données de l'ensemble des corpus ne fournit d'ailleurs aucun significativité (sous R et avec un modèle simple, p = 0,21 pour CFPP et p = 0,23 pour l'ensemble des corpus). Cependant, les données suggèrent nettement que *bon* est plus fortement associé à *mais* que ne l'est *d'ailleurs*.

Quatre approches (au moins) permettent d'affiner l'analyse. Premièrement, lorsqu'on se limite aux corpus oraux, on trouve une différence significative (p = 0,016). D'une manière générale, nous tenons compte de la variation linguistique en calculant les contrastes de ce type pour tous les sous-ensemble de corpus et certains sous-ensembles de mesures. Par ailleurs, lorsque nous disposons de données démographiques (âge, sexe, catégorie socio-professionnelle, etc.), nous les intégrons comme facteur en interaction (ou comme facteur aléatoire dans un modèle hiérarchique). Deuxièmement, nous utilisons une distance entre vecteurs. Par exemple, pour les vecteurs de valeurs sur l'ensemble des corpus, la similarité cosinus[1] donne un cosinus de 0.53, ce qui correspond à un angle de 58°, donc à une distance prononcée. Le plus intéressant est d'avoir un éventail d'associés d'un même pivot MD comparés entre eux à un pivot. Par exemple, en comparant les vecteurs de *bon*, *alors*, *quand même*, *d'ailleurs*, *enfin* et *du coup*, on constate (figure 2) que les deux premiers sont les plus proches (13,8°).

1 . Pour deux vecteurs numériques v1 et v2, simcos(v1,v2) = _____v⟨1v1,v2v⟩2 , le quotient du produit scalaire par le produit || ||×|| || des normes.

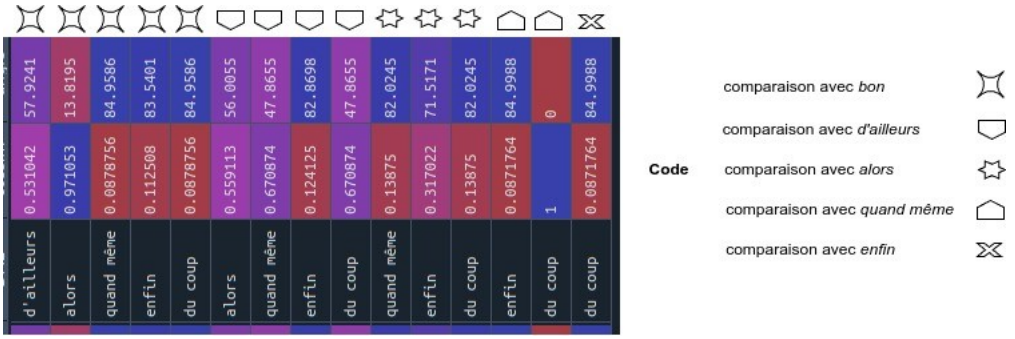


figure . 2 Similarité cosinus pour six MD sur l'ensemble des corpus

Troisièmement, pour éviter l'instabilité due aux estimations numériques, on peut « binariser » les résultats. Par exemple, pour les données de la figure 1, on va utiliser deux valeurs complémentaires (1/0 typiquement) pour noter le fait que la valeur pour *bon* est supérieure ou égale (1) ou inférieure (0) à la valeur pour *d'ailleurs* pour la même mesure. Une régression logistique (Hosmet et al. 2013) montre alors que la différence est très significative ($p \approx 0$). Quatrièmement, on peut normaliser les valeurs pour les cantonner à l'intervalle [0,1].

Parallèlement à ces méthodes, nous utilisons des outils d'agrégation (*clustering*). Le scénario le plus simple consiste, comme pour la figure 2, à choisir un pivot et à construire les vecteurs d'associations pour différents MD. Les scores obtenus (angle, résultat d'une binarisation) peuvent alors être comparés et regroupés en classes au moyen de différents algorithmes d'agrégation. Nous testons également l'apport des analyses en composantes principales (ACP) dans des configurations où les MD jouent le rôle des « individus » et les mesures celui des propriétés.

Développements

Les deux principaux développements en cours ou à venir concernent.

(a) La question de la composition des associations. Par exemple, faut-il considérer que *mais alors bon* résulte d'une coagulation entre *mais* et *alors* suivie de *bon* (motif noté [*mais alors*][*bon*]), ou plutôt d'un motif [*mais*][*alors bon*], ou d'un enchaînement de deux motifs [*mais alors*] et [*alors bon*]?

(b) Les répétitions et les distances. Nous analyserons deux images alternatives des données présentéesà la section 3 : une image où on contingente les répétitions et où on supprime les marques d'hésitation (*euh* et variantes) et une image où on tient compte des MD non contigus, en limitant la fenêtre droite, ce qui est nécessaire pour les connecteurs, souvent présents après l'auxiliaire (*a*/*est donc*/*pourtant*/*encore*, etc.), et pour les MD à droite des unités prédicatives (*quoi*, *hein*, *d'accord*, etc.).

Références bibliographiques

Anscombre, J.C., Donaire, M.L., Haillet, P.P. (Éds) (2013). *Opérateurs discursifs du français*. Berne :Peter Lang.

Anscombre, J.C., Donaire, M.L., Haillet, P.P. (Éds) (2018). *Opérateurs discursifs du français, 2*. Berne :Peter Lang.

Brezina, V. *Statistics in Corpus Linguistics. A Practical Guide*. Cambridge : Cambridge University Press.

Crible, L., Degand, L. (2021). Co-occurrence and ordering of discourse markers in sequences : A multifactorial study in spoken French. *Journal of Pragmatics*, 177, 18-28.

Crible, L., Degand, L. (2024). Discourse markers and quantitative corpus linguistics. Dans : Mosegaard Hansen, M.-B., Visconti, J. (Éds.), *Manual of Discourse Markers in Romance*, Berlin/Boston : de Gruyter, 257-279. Cuenca, M.-J. (2024). Clusters of discourse markers. Dans : Mosegaard Hansen, M.-B., Visconti, J. (Éds.), *Manual of Discourse Markers in Romance*, Berlin/Boston : de Gruyter, 193-223.

Dargnat, M. (2022). *Mais enfin* : construction et association. *Langages*, 225, 49-63.

Dargnat, M. (2024). Les particules énonciatives, in *Encyclopédie grammaticale du français*, en ligne : encyclogram.fr (http ://www.encyclogram.fr/notx/030/030_Notice.php)

Dostie, G. (2004). Les associations de marqueurs discursifs. De la cooccurrence libre à la collocation. *Linguistik Online*, 62, 5-13.

Evert, S. (2005). *The Statistics of Word Cooccurrences Word Pairs and Collocations*. Thèse de doctorat. Institut für maschinelle Sprachverarbeitung, Université de Stuttgart.

Haselow, A. (2018). Discourse marker sequences : Insights into the serial order of communicative tasks in real-time turn production. Journal of Pragmatics, 146, 1-18.

Hosmer, D. W., Jr, Lemeshow, S., Sturdivant, R. X. (2013). *Applied Logistic Regression*. Hoboken : Wiley.

Jayez, J. (2025). Discourse markers are not special (but they can be complicated). In Bîlbîie, G., Schaden, G. (Éds.), *Empirical issues in syntax and semantics : Selected papers from CSSP 2023*, 171196. Berlin : Language Science Press (https ://zenodo.org/records/15450440/files/519-BilbiieSchaden-2025-7.pdf?download=1).

Koev, T. (2019). *Parenthetical Meaning.* Oxford : Oxford University Press.

Koops, C., Lohmann, A. (2015). A quantitative approach to the grammaticalization of discourse markers : evidence from their sequencing behavior. *International Journal of Corpus Linguistics*, 20.2, 232-259.

Lohmann, A., Koops, C. (2016). Aspects of discourse marker sequencing. Dans : Kaltenböck, G., Evelien, K., Lohmann, A. (Éds.), *Outside the Clause : Form and Function of Extra-clausal Constituents* (Studies in Language Companion Series v. 178). Amsterdam/Philadelphia John Benjamins Publishing Company, 417-446.

Roze,C. (2013). *Vers une algèbre des relations de discours.* Thèse de doctorat. Université Paris-Diderot.

Schneider, U. (2020). ΔP as a measure of collocation strength. *International Journal of Corpus Linguistics*, 16.2, 249-274.

Les arrêts de la Cour Suprême des Etats-Unis entre narration, autorité et argumentation : une analyse discursive et linguistique sur corpus

Warren Bonnard [1]
[1] Laboratoire ATILF, Université de Lorraine & CNRS
warren.bonnard@univ-lorraine.fr

Introduction

Les opinions judiciaires, en particulier celles écrites par la Cour suprême des États-Unis (SCOTUS), constituent un genre important pour l'enseignement de l'anglais juridique (*English for Legal Purposes* – ELP). Ces textes, souvent dépourvus de structure explicite et fortement marqués par des variations stylistiques (Yaich & Hernandez, 2025), posent des difficultés de compréhension aux apprenants d'ELP, peu familiers des conventions rhétoriques de la common law (Kirby-Légier, 2005). Comprendre ces textes implique non seulement l'acquisition d'un nouveau mode de raisonnement juridique, mais aussi la maîtrise de conventions linguistiques et discursives propres à cette tradition.

Les travaux en analyse de genre (Swales, 1990 ; Bhatia, 1993) ont montré que l'identification des fonctions communicatives récurrentes – ou *moves* – permet aux apprenants de langue seconde de mieux appréhender la structure des textes spécialisés. Toutefois, la correspondance entre ces fonctions et leurs réalisations linguistiques reste souvent floue (cf. la notion de « *form-function gap* » ; Moreno & Swales, 2018). Pour répondre à cette difficulté, certaines études (Kanoksilapatham, 2008 ; Gray et al., 2020) ont mis en œuvre une approche combinant analyse de genre et caractérisation linguistique au moyen de méthodes quantitatives.

Dans cette perspective, la présente étude applique l'analyse de genre à un corpus d'opinions majoritaires de la SCOTUS. Les fonctions communicatives identifiées sont ensuite groupées en unités de discours cohérentes, dont les textes associés font l'objet d'une analyse multidimensionnelle (MDA ; Biber, 1988). Cette approche permet d'identifier, dans un corpus textuel, les cooccurrences de variables linguistiques et de les interpréter en termes fonctionnels. L'objectif de l'étude est ainsi de mettre en évidence les liens entre fonction communicative et forme linguistique, afin d'élaborer des stratégies pédagogiques adaptées à l'enseignement de la lecture de ces textes en anglais juridique.

Corpus et méthodologie

Corpus

Le corpus comprend 256 arrêts majoritaires de la SCOTUS, rédigés entre 1945 et 2020. Il est divisé en deux sous-corpus complémentaires :

- NORM : 180 arrêts représentatifs des décisions ordinaires, échantillonnés afin de refléter la diversité des thèmes et des auteurs des arrêts de la SCOTUS ;

- LAND : 76 décisions dites *landmark,* sélectionnées pour leur impact sociétal majeur, à partir du critère de leur fréquence de citation dans les revues universitaires juridiques.

Ce corpus a ensuite été annoté selon les méthodes de l'analyse de genre (Biber et al., 2007, Moreno & Swales, 2018), afin de décrire les fonctions rhétoriques (*moves*) spécifiques aux arrêts SCOTUS. Un schéma d'annotation a été élaboré de manière inductive en prenant la phrase comme unité minimale. Faute de pouvoir identifier des moves, soit des unités de discours stables et régulièrement réalisées par des fonctions communicatives de niveau inférieur, le schéma propose quatre catégories de fonctions communicatives à la portée plus large, apparaissant dans un ordre relativement fixe : Mise en Scène, Sources d'Autorité, Analyse, et Résolution.

Pour les besoins de l'analyse multidimensionnelle, une unité discursive est définie comme une séquence de phrases consécutives annotées avec la même catégorie. Seules les unités de plus de 100 mots ont été retenues pour l'analyse linguistique, soit 1 816 unités exploitables.

Catégorie	Total	LAND	NORM
Mise en Scène	462	151	311
Sources d'Autorité	303	129	174
Analyse	1 014	414	600
Résolution	37	21	16
Total général	1 816	715	1 101

table 1. : Distribution des unités discursives (≥ 100 mots) par sous-corpus et par catégorie

Méthodologie

Chaque texte relatif à une unité de discours a ensuite été annoté avec le BiberTagger (Biber, 1988), qui permet de calculer, par texte, des fréquences normalisées de certaines variables lexico-grammaticales. Une sélection de 96 variables liées à la persuasion, à la narration, à la densité informationnelle ou à la complexité grammaticale a été testée, puis réduite à 31. Celles-ci incluent par exemple : la longueur moyenne des mots, des catégories sémantiques de verbes (cognitifs, communicationnels, modaux), les pronoms, ou encore des structures syntaxiques (complétives avec *that*, subordonnées adverbiales, etc.).

Une analyse en composantes principales (PCA) a mis en évidence des facteurs de co-occurrence entre les variables linguistiques choisies pour l'analyse. Une solution à 4 facteurs a été retenue, expliquant 34 % de la variance totale. Les variables linguistiques en co-occurrence dans chaque facteur ont ensuite permis d'interpréter chaque facteur en termes de dimension de variation linguistique dans le corpus étudié.

Dimension	Variables aux charges positives les plus élevées dans la dimension	Variables aux charges négatives les plus élevées dans la dimension
1. Implication et évaluation	Verbes au présent, modaux, adverbes, adjectifs évaluatifs	Verbes au passé, noms propres, autres verbes de communication
2.Raisonnement interne	Verbes mentaux, pronoms 1re pers., complétives avec *that* évaluatives	Conjonctions
3.Densité informationnelle	Nominalisations, diversité lexicale, longueur des mots	Articles définis
4.Cohérence logique	Subordonnées adverbiales, pronoms 3e pers., conjonctions	Prépositions, articles définis

table 2. : Dimensions de variation linguistique et liste de caractéristiques lexico-grammaticales associées

Résultats

Chaque unité textuelle a reçu un score sur chaque dimension, en fonction de la présence au sein de l'unité de variables linguistiques à charge positive ou négative. Les interprétations des scores ont été affinées par l'analyse qualitative d'unités de discours présentant les valeurs les plus extrêmes. Ainsi, les unités Analyse se distinguent en général par des scores positifs élevés sur les dimensions 1 et 2, reflétant une forte implication discursive, une activité cognitive marquée, et une argumentation interprétative. Les unités Résolution, plus courtes et contribuant plus faiblement à l'extraction de facteurs du fait de leur faible nombre dans le corpus, présentent des profils similaires (cf. Figure 1). Cela peut s'expliquer par une concentration de conclusions interprétatives dans les dernières phrases des opinions.

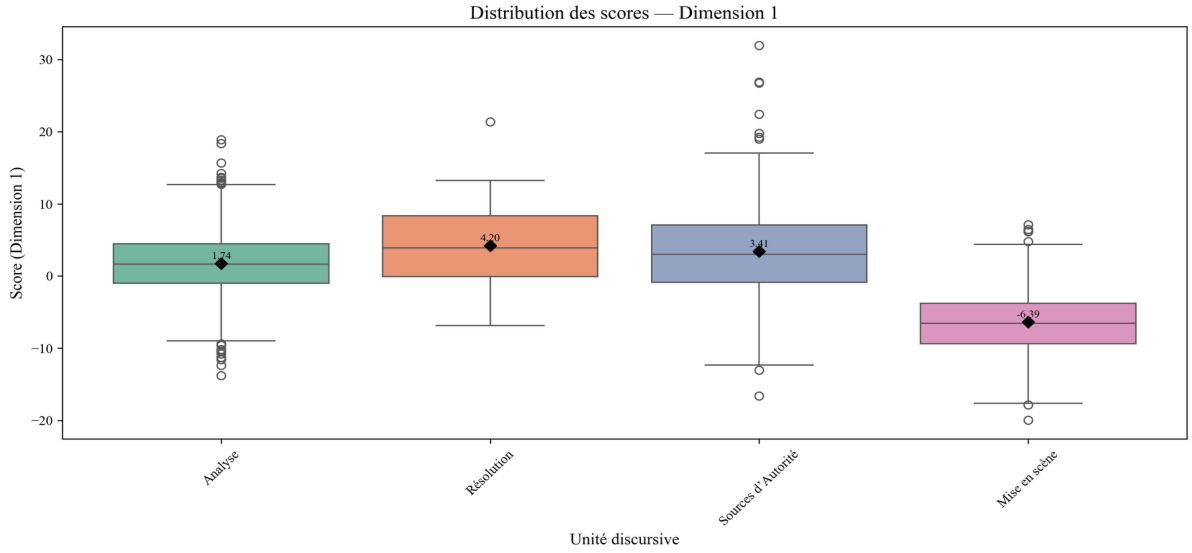


figure . 1 Distribution des scores des différentes catégories d'unités discursives sur la dimension 1

Les segments Mise en scène sont plus factuels et narratifs, caractérisés par des temps passés, des noms propres, et une faible présence de marqueurs évaluatifs. Ils obtiennent des scores faibles sur les dimensions 1 et 2, mais modérément élevés sur la densité lexicale (dimension 3). Enfin, les segments Sources d'autorité présentent la plus grande hétérogénéité, oscillant entre un style dense et technique (citations législatives) et un style administratif plus simple. Leur faible score en raisonnement interne confirme leur fonction de citation et non d'interprétation.

Des tests post-hoc ont par ailleurs été réalisés pour mettre en évidence l'effet de variables contextuelles dans la distribution des textes sur le continuum des dimensions. En particulier, l'année de décision et la thématique juridique influencent de façon répétée la distribution des traits linguistiques, notamment dans les segments Analyse et Mise en Scène. L'effet du corpus (LAND vs. NORM) est également significatif pour certaines dimensions et catégories discursives, ce qui suggère des différences de style entre arrêts ordinaires et décisions historiques. Toutefois, l'effet des juges, bien que modélisé, n'est pas isolé dans la majorité des cas, faute d'observations suffisantes par auteur.

Références bibliographiques

Bhatia, V. K. (1993). Analysing genre: Language use in professional settings. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9781315844992/analysing-genre-bhatia

Biber, D. (1988). Variation across speech and writing. Cambridge University Press.

Biber, D., Connor, U., & Upton, T. A. (2007). Discourse on the Move: Using corpus analysis to describe discourse structure (Vol. 28). John Benjamins Publishing Company. https://doi.org/10.1075/scl.28

Gray, B., Cotos, E., & Smith, J. (2020). Combining rhetorical move analysis with multi-dimensional analysis: Research writing across disciplines. In U. Römer, V. Cortes, & E. Friginal (Éds.), Studies in Corpus Linguistics (Vol. 95, p. 137-168). John Benjamins Publishing Company. https://doi.org/10.1075/scl.95.06gra

Kanoksilapatham, B. (2008). Rhetorical moves in biochemistry research articles. In Discourse on the move : Using corpus analysis to describe discourse structure (p. 73-119). John Benjamins Publishing Company.

Kirby-Légier, C. (2005). Understanding the judicial discourse of the current United States Supreme Court. La Langue, le discours et la culture en anglais du droit. Paris : Publications de la Sorbonne, 87-110.

Moreno, A. I., & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. English for Specific Purposes, 50, 40-63. https://doi.org/10.1016/j.esp.2017.11.006

Swales, J. M. (1990). Genre analysis. Cambridge University Press.

Yaich, M., & Hernandez, N. (2025). Improving Accessibility of SCOTUS Opinions : A Benchmark Study and a New Dataset for Generic Heading Prediction and Specific Heading Generation.

Les emplois du marqueur *par exemple* dans les interrogatives des enquêteurs du *Discours sur la ville*

Corinne Gomila
PRAXILING, Université de Montpellier Paul-Valéry
corinne.gomila@univ-montp3.fr

Introduction

Si l'exemplification est appréhendée le plus souvent au détour d'études portant sur les marqueurs discursifs, plusieurs travaux (Dostie 2002 ; Rossari et Jayez, 2003 ; Landolsi, 2018) dont un numéro récent de *Langue Française* coordonné par H. Landolsi et D. Vigier (2024) l'envisagent spécifiquement. Dans l'ensemble, tous s'entendent sur une description de base qui définit l'exemplification comme « une relation entre deux séquences X et Y : une séquence exemplifiée et une séquence exemplifiante, généralement introduite par un marqueur ». *Par exemple* est reconnu comme le marqueur prototypique de cette relation.

Soit l'extrait suivant :

[1] On proposa aux photographes de travailler sur des thèmes, l'auto-portrait par exemple, parce que l'appareil se prête bien à cet exercice solitaire […] (Frantext : Guibert Hervé – 1981)

En [1], selon les analyses d'H. Landolsi (2018) notamment, cette relation se présente sous la forme d'une relation d'inclusion/appartenance qui rattache l'exemple introduit par le marqueur – ici *l'auto-portrait* - à une classe paradigmatique d'éléments – *des thèmes* - mentionnée ou pas. De fait, les éléments de cette classe partagent des points communs que partage également l'exemple qui en est extrait.

Notre propos vise à questionner les emplois du marqueur d'exemplification *par exemple* au sein d'un corpus oral, le *Discours sur la ville,* Corpus du Français Parlé Parisien des années 2000[1]. Il a été choisi pour son genre – l'entretien - propice à l'exemplification et le fait que *par exemple* en termes de fréquence, y arrive en tête de liste des adverbes ou locutions adverbiales susceptibles d'introduire un exemple.

Corpus et méthodologie

Corpus

Initié par S. Branca-Rosoff, le Corpus du Français Parlé Parisien des années 2000 (désormais CFPP2000) est une base de données sur le français parlé de la région Ile-de-France. Il a été

1 Ce corpus est disponible sur le site de l'Université de Paris 3 (http://cfpp2000.univ-paris3.fr/ . Il a été recueilli par S. Branca-Rosoff (Université Sorbonne Nouvelle), F. Lefeuvre (Université Sorbonne Nouvelle) et M. Pires (Université de Besançon). L'organisation et la mise en ligne des données sont assurées par S. Fleury (Université Sorbonne Nouvelle).

constitué à partir d'« entretiens longs, semi-directifs qui portent sur le rapport des Parisiens à leur quartier » (Branca-Rosoff, Fleury, Lefeuvre, Pires, 2012). Plusieurs thématiques sont abordées lors de l'entretien. Les habitants sont questionnés par exemple sur les changements survenus dans leur quartier au fil du temps, leurs modes de vie ou encore les pratiques plurilingues qu'ils partagent ou qui les entourent. Ces demandes suscitent de longues prises de parole, sous des formes variées, selon qu'elles sollicitent des récits, des descriptions ou des justifications.

Disponible en ligne (http://cfpp2000.univ-paris3.fr/Corpus.html), le corpus rassemble à ce jour 61 entretiens. Il totalise une durée d'enregistrement de plus de 72 heures et un nombre d'occurrences qui approche le million (989 058 occurrences). Plusieurs données sont accessibles sur le site : ainsi les fichiers de transcriptions et les fichiers audios associés permettent à partir du logiciel Transcriber ou via Cocoon de disposer d'une lecture alignée de chaque entretien. Il est ainsi possible d'écouter une interview tout en lisant sa transcription structurée en tours de parole. Cette option nous a permis notamment de résoudre des problèmes de délimitation d'énoncés et de portée du marqueur *par exemple*. L'ensemble des données du corpus est également associé à des métadonnées documentant entre autres le profil des locuteurs (âge, sexe, profession, etc.), enquêteurs comme enquêtés.

Il est possible de lancer des requêtes sur les contenus des données de transcription du CFPP2000. Différents modes recherches sont proposés comme une recherche de mots/séquences de mots ou une recherche du motif questionné via *la carte des sections* de l'entretien (fêtes, transports, division de l'espace parisien, etc.). Enfin, le corpus est disponible avec ses propres outils d'interrogation, notamment l'outil textométrique *iTrameur*. Il est possible d'importer directement la « base CFPP2000 » (32 entretiens cette fois-ci/ 662687 occurrences/ 17711 formes) comportant 3 couches d'annotations (forme, lemme, catégorie) un découpage en tour de parole et 3 systèmes de parties (par quartier, par transcription et par interlocuteur). Les données peuvent être alors questionnées via les opérations classiques : dictionnaire, concordancier, spécificités, segments répétés, cooccurrents, etc.

Méthodologie

Les 621 occurrences du marqueur *par exemple* ont été extraites directement de la base de données CFPP2000. Dans un premier temps, elles ont été triées et questionnées en contexte selon le statut des locuteurs qui les emploient, enquêteurs ou enquêtés. Dans un second temps, au sein du sous corpus rassemblant les occurrences de *par exemple* présentes dans les interrogatives des enquêteurs, les propriétés syntaxique, sémantique et pragmatique du marqueur seront analysées.

Résultats

Une première exploration du corpus montre que la locution adverbiale est présente dans toutes les interviews, avec une moyenne de 15 emplois par interview. Comme on peut s'y attendre, les enquêtés l'utilisent régulièrement pour exemplifier leur propos. Elle fonctionne comme un marqueur d'exemplification dans trois types d'emploi (Landolsi, 2018) :

Ainsi, en [2], l'exemple introduit par le marqueur est un constituant facultatif de la proposition. Il illustre la classe à laquelle il appartient et dont il est un *échantillon* ; en [3], l'exemple est cette fois-ci un constituant obligatoire qui se trouve doté d'une certaine singularité, ce qui en fait un exemple à part, *un modèle* du genre ; en [4], il est une proposition entière, un *argument* venant confirmer l'énoncé précédent, et qui, en le renforçant, le rend plus réel.

[2] spk2 [2625.251] : d'action éducative voilà [c'est ; 0] ça et euh donc c'est les profs en fait qui prennent en charge ça qui créent un projet par exemple partir à Prague[2] […] (CFPP2000-07-04)

[3] spk2 [1036.129] : faudrait dé- faudrait imaginer un péril extérieur contre lequel se fédèrent des gens [mm] qui sont par ailleurs + profondément individualistes alors on pourrait imaginer + euh + l'imposition d'un camp de réfugiés *par exemple* [mm mm] là j'imagine qu'un certain nombre de gens dans le quartier […] euh se fédèreraient (CFPP2000-17-01)

[4] spk1 [550.234] : il y a des quartiers où vous ne seriez pas allés

spk2 [553.171] : non pas des quartiers ou on ne serait pas allés mais + un éloignement de métro ou + de commerçants

spk3 [560.206] : voilà + *par exemple* on a visité un appartement moi qui me plaisait énorme + énormément mais qui était euh Porte d'Asnière + très loin de tout euh + l'appartement était magnifique moi il me plaisait énormément il était très très beau et + […] euh + mais + il était loin pas de commerçants pas de bus euh + donc euh (CFPP2000-05-01)

Plus inattendue est la fréquence d'emploi de *par exemple* chez les enquêteurs. Le marqueur apparait quasi exclusivement dans des interrogatives variées - intonatives, averbales, en *est-ce que,* interro-négatives. En [5], la relation d'exemplification se construit en trois temps dans la co-énonciation. L'enquêteur revient sur l'énoncé de son interlocuteur. Il sélectionne via l'interrogative averbale *quoi par exemple* une classe à exemplifier. Le marqueur n'introduit pas d'exemple directement. C'est à l'enquêté de s'en charger.

Mais le plus souvent, pour poser leurs questions, les enquêteurs exemplifient leurs propres demandes. En [6], l'acte d'exemplification s'actualise dans l'interro-négative sous la forme d'un exemple focalisé qui peut être exploité ou nié, mais en aucun cas rester sans suite, ce qu'a très bien compris l'enquêté qui développe sa réponse. En [7], l'exemple introduit n'est pas le seul à illustrer la classe. Cette mise en série minimise la représentativité de l'exemple, laissant l'interlocuteur libre de poursuivre sur des exemplaires *équivalents.* Soulignons enfin, en [8], que le marqueur, fonctionnant avec le sens d'un « et pour parler d'autres choses » se désémantise. Il n'introduit pas d'exemple mais initie un nouveau thème dans l'échange (Sitri, 2003, 69).

[5] spk2 [1344.452] : […]euh + finalement si c'est de saison euh c'est enfin ça vient pas de l'autre bout du monde donc y a un coût en moins et y a y a quand même certains produits qui sont moins cher euh (mm)

spk1 [1366.667] : quoi par exemple […]

spk2 [1370.14] : non mais par exemple vous achetez des carottes ça coûte rien

[6] spk1 [1857.587] : d'accord donc euh à la limite c'est pas un soucis les par exemple le je sais pas le chômage c'est pas une peur ou euh

spk2 [1865.798] : euh non je fais tour pour je fais tout pour que ça ne le soit pas on va dire

2 Le soulignement est de notre fait. Il délimite le groupe rythmique dont fait partie le marqueur.

[7] spk1 [2399.579] : euh est-ce que est-ce qu'il y a par exemple une euh une influence je ne sais pas des radios jeunes du verlan ou de ou pas du tout

spk2 [2408.76] : ils sont quand même petits (CFPP2000-07-06)

[8] spk2 [1595.858] : voilà [pause] les jeux vidéo aussi j'aime bien […]et la musique aussi voilà

spk1 [1603.809] : d'accord [pause] et par exemple [pause] comme vous me disiez que vos parents venaient euh d'Algérie [pause] il y avait pas une communauté de gens qui venaient aussi d'Algérie que vous retrouviez

L'acte d'exemplification sert la conduite de l'entretien. Dans les interrogatives de l'enquêteur, l'exemplification participe du *faire dire.* Tout en conservant sa fonction illustrative, l'exemple introduit est un point d'appui offert sur lequel l'interlocuteur peut enchainer. De fait comme le montre le fonctionnement du marqueur en cooccurrence avec la conjonction *ou*, le phatique *euh* souvent couplé à un *je ne sais pas*, l'exemple introduit – au hasard ? - est régulièrement multiple [7], alternatif ou laissé en suspens [6].

Prenant appui en soubassement sur les recherches dédiées aux marqueurs discursifs (Dostie 2004 ; Dostie et Pusch 2007 ; Rouanne et Anscombre, 2016), aux marqueurs de reformulation (Rossari, 1990 ; Steuckardt, 2018 ; Fuch, 2020) et de glose (Steuckardt et Nikals-Salminen, 2003 ; 2005), l'étude s'adosse principalement aux travaux portant sur le marqueur *par exemple* et aux analyses consacrées à la variété des interrogatives (Borillo, 1979 ; Dagnac et Cappeau, 2013 ; Branca-Rosoff et Lefeuvre, 2017 ; Larrivée et Guryev, 2021).

L'analyse portera dans un premier temps sur les types d'interrogatives utilisées par les enquêteurs. Quelles sont les formes privilégiées ? quelle est la valeur ajoutée de la présence du marqueur dans l'interrogative qui pourrait tout à autant fonctionner sans. Dans un second temps, pour les interrogatives les plus représentées, nous procèderons à une description du fonctionnement du marqueur *par exemple* en analysant ses caractéristiques sémantique, syntaxique et pragmatique, notamment ses effets sur l'interlocuteur, avec une attention particulière prêtée à sa position et à sa portée dans l'énoncé. Nous nous demanderons si ces interrogatives "exemplifiées et exemplifiantes" mettent au jour des variantes d'emploi propres à l'acte d'exemplification, ou si, à l'instar de l'extrait [8], elles actualisent de nouvelles fonctions discursives pour le marqueur.

Références bibliographiques

Borillo, A. (1979). La négation et l'orientation de la demande de confirmation. *Langue française,* 44, 27-41.

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

Branca-Rosoff, S., Lefeuvre, F. (2016). Le Corpus de Français Parlé Parisien des années 2000 : constitution, outils et analyses. Le cas des interrogatives indirectes. In M. Avanzi, M.-J. Béguelin, F. Diémoz (éds). Corpus de français parlé et français parlé de corpus*, Cahiers Corpus,*15, 265-284.

Dagnac, A., Cappeau, P. (2021). La variation dans les phrases interrogatives. In A. Abeillé, D. Godard, *La Grande Grammaire du Français,* Actes Sud, 1432-1437.

Dostie, G. (2002). L'exemplification de *par exemple*. Un cas de pragmaticalisation en français québécois. *Journal en French Language Studies,*12, 149-167.

Dostie, G. (2004). *Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique.* Bruxelles, De Booeck-Duculo.

Dostie, G., Pusch, C. (2007). Présentation. Les marqueurs discursifs. Sens et variation. *Langue française,*154, 3-12.

Fuch, C. (2020). Paraphrase et reformulation : un chassé-croisé entre deux notions. *Recherche & Rencontre*, 36, 41-55.

Landolsi, H., Vigier, D. (2024). L'exemplification en français : perspectives linguistiques. *Langue française,* 222.

Landolsi, H. (2018). L'exemplification et ses marqueurs. Uppsala, Acta Universitatis Upsaliensis. *Studia Romanica Upsaliensa.* 86. 312 p.

Larrivée, P., Guryev, A. (2021) Variantes formelles de l'interrogation. Présentation. *Langue française,* 212, 9-24.

Rossari,C., Jayez, J. (2003). *Par exemple* : une procédure d'exemplification par la preuve. In B. Combettes, C. Schnedecker, A, Theissen (éds), *Ordre et distinction dans la langue et le discours.* Paris, Honoré Champion, 461-478.

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

Rouanne, L., Anscombre, J.-C. (2016). *Histoire de « dire » : petit glossaire des marqueurs formés sur le verbe « dire ».* Berne, Peter Lang.

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

Steuckardt, A., Niklas-Salminem, A. (2003). *Le mot et sa glose.* Aix-en-Provence, Publication de l'Université de Provence.

Steuckardt, A., Niklas-Salminem, A. (2005). *Les marqueurs de glose.* Aix-en-Provence, Publication de l'Université de Provence.

Sitri, F. (2003). *L'objet du débat. La construction des objets de discours dans des situations argumentatives orales.* Paris. Presses de la Sorbonne Nouvelle.

Les marqueurs de genre dans les interactions de la fiction policière (1945-1989)

Camille Bouzereau [1], Dominique Longrée[2] et Adam Faci [3]
[1] Laboratoire CSLF, Université Paris Nanterre
[2]SeSLa, Université Saint-Louis Bruxelles
[3] Huma-Num Lab
camille.bouzereau@parisnanterre.fr, dominique.longree@uliege.be, adam.faci@huma-num.fr

Introduction

Notre communication s'inscrit dans un projet de recherche plus large qui étudie les variations linguistiques des discours oraux représentés dans le roman policier publié en France de 1945 à 1989. Cette recherche articule les études quantitative et qualitative, tout en s'inscrivant en textométrie littéraire, dans la lignée des travaux d'Etienne Brunet (2003)[1].

Corpus et méthodologie

Corpus

Le corpus étudié est composé de 3015 romans policiers publiés en France, durant la guerre froide[2]. L'oral occupe une place prépondérante dans la fiction policière à partir de la seconde moitié du XXe siècle car le polar français investit alors à la fois l'héritage du roman des bas-fonds né au XIXe et du roman prolétarien de l'entre-deux-guerres (Levet 2024). Dès lors, notre étude linguistique porte sur ces variations diachroniques, diastratiques et diagéniques qui figurent dans les dialogues de la fiction criminelle. La problématique de recherche peut-être définie comme suit :

Les dialogues représentés dans le roman policier de la seconde moitié du XXe siècle attestent-ils de variations diachroniques, diastratiques et diagéniques ?

Cette problématique entraîne un ensemble de sous-questions, telles que : comment sont représentés les discours des personnages déclassés ou issus des classes populaires dans le néo-polar au regard du roman noir ? quelle place le roman noir laisse-t-il à l'argot et le néo-polar au phénomène néologique ? comment s'écrivent les stéréotypies au féminin et au masculin à travers les discours représentés ? C'est à cette dernière question que notre article tentera de répondre.

Plus précisément, pour les JLC 2025, nous présenterons une étude consacrée aux variations diagéniques, c'est-à-dire à la place que laisse le roman policier au discours genré, dans un

1 Ce projet participe à l'axe de recherche « linguistique textuelle » du SeSLa « Séminaire des Sciences du langage » de l'UCLouvain – Site Saint-Louis Bruxelles, coordonné par Anne Dister et Dominique Longrée.

2 Ce corpus a été constitué pour l'ANR POLARisation : https://anr.fr/Projet-ANR-22-CE54-0008

contexte bien particulier, à savoir celui d'un passage où un personnage féminin se fait gifler par un personnage masculin. Cette recherche s'inscrit dans la lignée d'un chantier mené avec Laetitia Gonon (MCF, Université de Rouen) et d'Adam Faci (postdoctorant, HumaNum Lab) autour d'un sous-corpus composé de 5000 occurrences de la base *gifl** (e.g. le substantif *gifle* et le verbe *gifler*). Dans ces contextes, nous étudions linguistiquement la représentation des corps et des discours féminins via l'analyse des chaînes de référence.

Méthodologie

Diverses études antérieures ont montré qu'un des paramètres pouvant servir à caractériser des types de discours était régulièrement, dans divers genres (Novakova & Sipman, 2020; Legallois et alii, 2016, 2020), mais en particulier dans le roman policier (Gonon et alii, 2018), l'emploi de "motifs textuels". Ces derniers peuvent être définis comme des patrons linguistiques (association de n éléments du texte muni de sa structure linéaire) récurrents dotés de fonctions textuelles, principalement de fonctions structurantes (par exemple, transitionnelles entre deux épisodes) ou caractérisantes (d'un style, d'un mode de discours, d'un genre). Les patrons lexico-grammaticaux formant motifs ne sont la plupart du temps pas figés, mais présentent diverses variantes sur base de permutations, substitutions, suppressions, variations morphologiques ou lexicales parmi les éléments qui actualisent le schème sous-jacent au motif (ex. : *ici, je voudrais toutefois préciser / je souhaite néanmoins préciser ici / je voudrais maintenant souligner*, etc.) (Longrée & Mellet, 2008, 2013, 2018, Mellet & Longrée 2012) L'outil conceptuel que constitue la notion de "motif textuel" a déjà été appliqué avec succès pour différencier divers sous-genres dans le roman et nous explorerons ici son potentiel pour caractériser le discours féminin dans le cadre de "scènes de gifles" et l'opposer au discours masculin. Il s'agira tant d'identifier des motifs propres au discours féminins dans ce contexte que d'en étudier les diverses réalisations caractéristiques de la langue ou du style de chaque auteur ou groupe d'auteurs.

Résultats

Les critères formels de définition des "motifs textuels" en font des objets susceptibles d'être détectés, identifiés et traités par diverses méthodes d'analyse des données textuelles ou par le Deep Learning (Longrée & Vanni, 2025). Dans le cadre de ce travail, nous nous appuyons sur le travail d'Adam Faci (2022) pour repérer des motifs textuels présentant la structure suivante : *verbe de parole* + *gifl** + *circonstant pour* + *pronom*. Le retour au texte permet de montrer qu'en l'occurrence, lors des scènes de violence conjugale, les discours féminins n'occupent que très peu l'espace discursif et c'est le discours narrativisé qui est privilégié (« *elle avait prononcé des mots regrettables. Il avait dû la gifler pour la faire taire » ; « elles se mirent à glapir des injures et Hernandez en gifla une pour lui imposer silence* »). Au moment de la gifle, le discours narrativisé sert à clore le dialogue - puisqu'à ce moment du script, il s'agit de faire assimiler au lecteur que le dialogue n'est plus possible. C'est l'effet escompté par les gifles : en giflant, le personnage masculin a pour finalité de faire taire le personnage féminin.

En tenant compte de cette caractéristique du roman policier, nous prolongerons l'étude en nous intéressant aux interactions présentes dans ces scènes de gifles telles que celles présentées dans les quatre extraits suivants :

(1) Elle reçut une **GIFLE** qui lui secoua la tête, alors elle se durcit et le regarda avec colère. — **Tu n'as pas le droit !** (Y a pas de bon Dieu, Amila, Série Noire, 1950).

(2) **Elle cria : — Non !** Clac ! Clac ! Une paire de **GIFLES**, rien de plus. — Tu vas parler, salope ! — **Non !** Clac ! Clac ! **— Au secours ! A l'assassin !** (*OSS 117 n'est pas mort*, Bruce, Mystère, 1953).

(3) Pour la dernière fois, sortez de ce lit. **— Pas je vouloir...** Elle se jeta sur le drap, tentant de le remonter. Paire de **GIFLES**, et je lui jetai son collant, son slip en pleine figure. — Vous le faites vous-même ou je le fais. **— Sviňa, prasa, bracové.** (*Hôtel des vieux aigles*, Rank, Espionnage, 1980).

(4) **— Je ne… Je ne…** Elle pleurait. **UNE GIFLE. Un cri.** — Où sont les photos, Georgia ? **— Je ne…** Encore **UNE GIFLE. Un nouveau cri. Des sanglots.** La voix douce de l'homme : — Ça va devenir bien pire, bien pire, ma cocotte. Des sanglots. (*Scoop toujours*, Peterson, Série Noire, 1989)

Avant de nous focaliser sur l'utilisation de "motifs textuels", dans un dernier temps, nous décrirons le rôle qu'ont ces dialogues dans la mise en place de l'intrigue. Nous tenterons ensuite de déterminer plus globalement des traits communs entre ces énoncés de personnages féminins en train de subir une agression. Dans les quatre exemples ci-dessus, on relève déjà quelques traits caractéristiques : sur-utilisation de marqueurs de négations, énoncés coupés sur le vif ((2) et (4)) ou encore énoncés syntaxiquement incorrects "pas je vouloir" (en 3). En outre, on y détecte une expression stéréotypée "tu n'as pas le droit" (en 1), récurrentes au sein du corpus dans les scènes de violence conjugale et constituant un motif caractérisant. Nous nous demanderons également comment, en opposition, s'écrit la stéréotypie au masculin à travers ces discours représentés.

Références bibliographiques

Brunet E. (2003). Peut-on mesurer la distance entre deux textes ?, La distance intertextuelle, Revue Corpus, n°2.

Gonon L, Goossens V, Kraif O., Novakova I. & Julie Sorba J. (2018), « Motifs textuels spécifiques au genre policier et à la littérature "blanche" », SHS Web Conf., 46 (2018) 06007, DOI: https://doi.org/10.1051/shsconf/20184606007

Faci A. (2022), Représentation, simulation et exploitation de connaissances dans le formalisme des graphes conceptuels, thèse de doctorat, soutenue à Sorbonne Université.

Levet N. (2024) « Du polar « amerloque » au roman de langue verte : enjeux génériques du style « roman noir » », Polar et styles d'époque, Styles du roman policier (dir. Dominique Rabaté, Vincent Berthelier, Marc Vervel), URL : http://www.fabula.org/colloques/document12580.php, page consultée le 19 January 2025.

Longrée D., Mellet S. & Luong X. (2008). « Les motifs : Un outil pour la caractérisation topologique des textes », in S. H. Bénédicte Pincemin (éd.) JADT. Lyon : Presses Universitaires de Lyon, 733-744. https://hal.science/hal-01364605

Longrée D. & Mellet S. (2013). « Le motif : Une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours », Langages 189, 1 : 65-79.

Longrée D. & Mellet S. (2018). « Towards a topological grammar of genres and styles: a way to combine paradigmatic quantitative analysis with a syntagmatic approach », The Grammar of Genres and Styles: From Discrete to Non-discrete Units. Berlin & Boston: De Gruyter Mouton, 140-63.

Longrée D. & Vanni L. (2025).« Identification des motifs textuels. Entre statistique et deep learning », Corpus [En ligne], 27 | 2025, mis en ligne le 13 mai 2025, consulté le 16 mai 2025. URL : http://journals.openedition.org/corpus/10326 ; DOI : https://doi.org/10.4000/13woj

Legallois D., Charnois T. & Poibeau T. (2016). « Repérer les clichés dans les romans sentimentaux grâce à la méthode des “motifs” », Lidil 53 : 95-117. https://doi.org/10.4000/lidil.3950

DOI : 10.4000/lidil.3950

Legallois D. & Koch S. (2020). « The Notion of Motif Where Disciplines Intersect: Folkloristics, Narrativity, Bioinformatics, Automatic Text Processing and Linguistics », in I. Novakova & D. Siepmann (éd.), Phraseology and Style in Subgenres of the Novel. A Synthesis of Corpus and Literary Perspectives. Cham : Palgrave Macmillan, 17-46.

Mellet S. & Longrée D. (2012). « Légitimité d’une unité textométrique : le motif », in A. Dister, D. Longrée & G. Purnelle (éd.), Actes des 11e Journées internationales d’Analyse statistiques des Données Textuelles - JADT 2012. Liège, 715–728.

Novakova I., & Siepmann D. (éd.). (2020). Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives. Cham : Palgrave Macmillan. DOI : 10.1007/978-3-030-23744-8

Les répétitions dans la parole des professionnels de santé
Des corpus différents pour des usages différents

Mylène Blasco[1] et Paul Cappeau[2]
[1] (UCA-LRL)
[2] (FoReLLIS)
Mylene.BLASCO-DULBECCO@uca.fr, paul.cappeau@univ-poitiers.fr

De nombreux travaux, dans des cadres théoriques variés, ont porté sur la répétition dans le discours et son rôle à l'écrit ou à l'oral (Paissa et Druetta (dir.) 2019 ; Prak-Derrington 2021). Mais le terme recouvre en fait divers phénomènes. La répétition peut être le fait d'un seul locuteur (auto-répétition) ou de plusieurs - généralement deux (hétéro-répétition). Dans les auto-répétitions, il peut s'agir d'une part d'un procédé de type rhétorique, qui est volontairement exploité par l'usager (locuteur ou scripteur). On l'observe par ailleurs dans l'oral spontané où elle relève du mode de production, non contrôlé, de ce médium (Blanche-Benveniste 1995 ; Krötsch 2007 ; Richard 2015, Ploog, 2015). Elle intervient, par exemple, dans les « réparations ». Enfin un troisième cas, que l'on observe notamment dans nos données (des consultations à l'hôpital), les répétitions reflètent des préoccupations des locuteurs et échappent en partie à son contrôle, sans pour autant être considérées comme des accidents. Elles participent d'une intention de communication. Les hétéro-répétitions, quant à elles, jouent un rôle important dans la gestion des échanges langagiers, tant du côté du producteur que du récepteur (Norick, 1987 ; Tannen,1989 ; Ursi et alii, 2018). Certaines fonctions ont été mises en lumière, notamment par les approches interactionnelles, autour de la question-réponse, de l'accord, de la négociation, de l'appropriation, de la focalisation etc. (Vion, 2004 ; André, 2011, Teston-Bonnard, 2017). La séparation entre ces divers emplois et certaines formes de reformulation ou certaines nuances à l'intérieur d'une même fonction et le lien avec des formes de reformulation s'avèrent parfois plus complexes à opérer dans les échanges attestés.

Le souci d'améliorer la communication avec le patient est exprimé (Adolphs et al. 2004, Weber, 2017 ; Blasco (éd) 2024) dans les conseils adressés aux professionnels de santé. Le savoir-faire du médecin, pour la relation médecin-patient, doit intégrer une écoute active. Celui-ci doit vérifier qu'il comprend ce que dit le patient mais il doit aussi se faire comprendre de lui. Ainsi, les médecins sont encouragés à *reformuler* en *répétant* les mots du patient. On peut s'interroger sur ce brouillage de la frontière entre *reformuler* et *répéter* dans les conseils adressés aux médecins. L'observation des faits de langue à travers des corpus permet d'éclairer cette question.

Notre étude s'inscrit dans la lignée des études syntaxiques sur corpus oraux. Elle convoque deux contextes interactionnels particuliers. Les échanges, à l'hôpital, entre un patient et un professionnel de santé prennent la forme d'une consultation (avec un médecin) ou d'une présentation clinique (avec un psychanalyste). Les faits de langue étudiés apportent des pistes de réflexion pour comprendre le fonctionnement de la répétition dans ces situations et saisir le

rôle qu'elle joue dans la construction de l'entretien. Ils obligent à creuser les interprétations pour montrer toute la finesse d'emploi du phénomène.

La première partie expose les précisions et les difficultés rencontrées pour l'étude des répétitions sur corpus. Plusieurs paramètres bien connus sont à prendre en compte comme la forme, la taille et la distance entre les segments répétés (Magri-Mourgues & Rabatel, 2015), mais, de plus, une attention particulière doit être accordée au contexte dans lequel sont collectées les répétitions. Ce sera l'occasion de signaler les problèmes de constitution du corpus.

La deuxième partie expose les spécificités du corpus d'entretiens à l'hôpital. Ce corpus se prête à une approche comparative entre, d'une part, des entretiens entre médecin et patient et, d'autre part, des entretiens entre un psychanalyste et un patient. Ces deux situations se différencient par le statut professionnel du locuteur et par la fonction qu'il occupe dans l'échange, par l'objectif de la rencontre et par les préoccupations qui sont au centre du contenu de l'échange. Ce contexte interactionnel ouvre des perspectives d'analyse (ici qualitatives) pour apprécier des stratégies communicatives et discursives. Les répétitions, dans leur fonctionnement, régulent de façon différenciée les échanges et jouent des rôles subtils dans le discours.

La troisième partie, centrée sur une analyse fouillée des données traite d'une part des auto-répétitions et d'autre part des hétéro-répétions. Les hypothèses fondées sur les faits de langue en discours montrent que, dans ces interactions, les répétitions participent de l'écoute nécessaire à la relation entre le patient et le professionnel de santé, mais médecins et psychanalystes s'emparent différemment de cette stratégie de communication. Elles jouent une part active dans l'échange, elles permettent de maintenir le dialogue en assurant des fonctions plutôt fines. Ces intentions peuvent être le souci de manifester une insistance dans différents buts, de maintenir une continuité thématique, de consigner une information, d'intégrer les mots de l'interlocuteur et leur évolution, ou encore de faire reposer son propre discours sur les mots de l'interlocuteur dans la perspective de co-construire un discours.

Chaque fonction peut être assorties de différentes nuances, le type d'entretien (CO vs PC) et le statut des locuteurs (MED-PSY-PAT) interviennent dans la reconnaissance des différents rôles discursifs et communicationnels. On peut penser que l'auto-répétition, qui s'inscrit dans le souci d'une écoute attentive, différencie la parole professionnelle de la parole individuelle et singulière. Les hétéro-répétitions quand elles s'inscrivent dans un contexte de questions-réponses servent les objectifs de la consultation mais elles sont aussi une stratégie du psychanalyste pour co-construire avec le patient un discours et encourager la parole libre.

Ainsi disposer, dans le cadre de la relation de soin, de deux situations distinctes apporte un éclairage singulier sur les formes et les fonctions des répétitions en mettant en avant le poids des usagers.

Quelques références bibliographiques

Adolphs, S., Brown, B. Carter, R. Crawford, C. et Sahota, O. 2004. Applying Corpus Linguistics in a Health Care Context. *Journal of Applied Linguistics*. 1. 9-28.

André, Virginie. 2011. « Analyse pragmalinguistique de la co-construction du discours : le cas des énonciations conjointes et des reprises ». *Actes du colloque XII International Symposium on Social Communication*, 17-21 janvier 2011, Santiago de Cuba, p. 322-326.

Anquetil, Sophie et Lefebvre-Scodeller, Sophie (dir.). (2020). « Dis-moi ce que tu répètes, je te dirai qui tu es ». *Espaces Linguistiques* N° 1. Université de Limoges : PUL.

Blanche-Benveniste, Claire. (1987). « Syntaxe, choix du lexique et lieux de bafouillage », DRLAV, n°36-37, 123-157.

Blanche-Benveniste, Claire. 1995. Répéter ou ne pas répéter. *Tendances récentes en linguistique française et générale*. Volume dédié à David Gaatone. Bat-Zeev Shyldkrot H., Kupferman L. (éds). Amsterdam-Philadelphia. Benjamins. 55-74.

Blanche-Benveniste Cl. (2003). La naissance des syntagmes dans les hésitations et répétitions du parler. J.-L. Araoui (Éd), Le sens et la mesure. Hommages à Benoît de Cornulier. Paris : Honoré Champion, 40-55.

Blasco, Mylène (éd). 2022. *Parler à l'hôpital. Écouter ce qui est dit, décrypter ce qui se dit*, Münster, Nodus Publikationen.

Mylène Blasco, Paul Cappeau. Ce que les phraséologismes nous apprennent sur la parole en contexte de soins. Etude de quelques constructions avec « dire ». *Lingvisticae investigationes : International Journal of Linguistics and Language*, 2023, 45 (2), pp.197-217.

Blasco, Mylène (éd). 2024. *Langage, langue, parole dans la relation de soin*. Clermont-Ferrrand. Presses universitaires Blaise Pascal.

Clinquart, Anne-Marie (2000). « La répétition, une figure de reformulation à réviser », in Anderson, Patrick, Chauvin-Vileno, Andrée, Madini, Mongi (coord.), Répétition, Altération, Reformulation : Colloque international 22-24 juin 1998, Besançon, PUFC, 2000, 323-349.

Henry, S. (2005). Quelles répétitions à l'oral ? Esquisse d'une typologie. G. Williams (Éd.), La Linguistique de corpus. Rennes : Presses universitaires de Rennes, 81-92.

Krötsch, Monique. 2007. Répétition et progression en français parlé. *LINX* 57. 37-46.

Magri-Mourgues, Véronique & Rabatel, Alain (2015). « Quand la répétition se fait figure », *Semen 38,*

Norrick N. 1987. « Functions of repetition in conversation". *Texte 7* (3). 245-264.

Paissa, Paola et Druetta, Rugerro. 2019. *La répétition en discours*. Paris. L'Harmattan.

Penn, Claire & Watermeyer, Jennifer. 2018. Communicating Across Cultures and Languages in the Health Care Setting. Voices of Care. London. Palgrave MacMillan.

Prak-Derrington, Emmanuelle. 2021. Magie de la répétition. Lyon. ENS.

Katja Ploog, Sophie Mariani-Rousset, Séverine Equoy Hutin. Emmêler & démêler la parole, Approche pluridisciplinaire de la relation de soin. Annales littéraires de l'université de Franche Comté. Presses universitaires de Franche Comté, 2018, 978-2-84867-634-0.

Élisabeth Richard, « A propos de répétition : entre continuité et rupture », Semen [Online], 38 | 2015, Online since 24 April 2015, connection on 19 February 2019. URL : http://journals.openedition.org/ semen/10323

Tannen, Deborah. 1989, *Talking voices. Repetition, dialogue, and imagery in conversational discourse*. Cambridge. Cambridge University Press.

Sandra Teston-Bonnard. Etude linguistique des hétéro-répétitions ‘‘à l’identique’’ (ou presque) dans des contextes interactionnels. *Repères-Dorif. Autour du français : langues, cultures et plurilinguisme*, 2017.

Ursi, B., Etienne, C., Oloff, F., Mondada, L. et Traverso, V. (2018). « Diversité des répétitions et des reformulations dans les interactions orales : défis analytiques et conception d’un outil de détection automatique ». *Langages* N° 212(4). 87-104. https://doi.org/10.3917/lang.212.0087.

Vion, Robert. 2006. « Reprise et modes d’implication énonciative ». *La linguistique*, vol. 42 : 11-28.

Weber, Jean-Christophe. 2017. *La consultation*. Paris. PUF.

Nouveaux outils pour le traitement et l'exploration de corpus multi-parallèles

Cyrille François [1], Elnaz Jalilian [2], Olivier Kraif [2] et Natacha Rimasson Fertin [3]
[1] CIEL, Université de Lausanne
[2] LIDILEM, Univ. Grenoble Alpes
[3] ILCEA4, Univ. Grenoble Alpes
cyrille.francois@unil.ch, elnazjalilian@gmail.com, olivier.kraif@univ-grenoble-alpes.fr,
natacha.rimasson-fertin@univ-grenoble-alpes.fr

Introduction

Les quinze dernières années, les outils numériques de d'alignement et de visualisation de corpus parallèles se sont largement développés dans le domaine des humanités numériques : on peut citer le logiciel HyperMachiavel (Gedzelman & Zancarini, 2011) pour un corpus de traductions du *Prince* de Machiavel, l'outil d'alignement du projet Odysseus de Marianne Reboul qui étudie les traductions de l'*Odyssée* en français du XVIe au XXe siècle (Reboul, 2022), le logiciel Collatex de Dekker & Middel (2011) qui permet d'étudier des alignements multiples au niveau lexical, ainsi que la plate-forme Variance (cf. https://variance.unil.ch) dédiée à la comparaison de différentes versions d'une même édition. Des outils existent depuis longtemps dans le domaine de la linguistique de corpus ou de la traductique, qu'il s'agisse d'aligneurs comme WinAlign (Moore, 2002), Hunalign (Varga et al., 2007), LF Aligner[1], JAM (Kraif, 2015[2]), ou de concordanciers parallèles tels que ParaConc[3], et de grands volumes de corpus parallèles sont consultables en ligne, notamment à travers la plate-forme OPUS de J. Tiedemann (2012) ou des outils plus grand public tels que Linguee ou Tradooit.

Malgré la maturité des techniques d'alignement et le grand nombre de projets menés dans le domaine, force est de constater que nombre de corpus parallèles élaborés dans un contexte académique ne sont pas (ou plus) consultables en ligne, faute de maintenance ou de finance pour développer des sites *ad hoc* (comme celui d'Hyperprince). Par exemple, le corpus parallèle CARMEL (El Bèze et al., 2006), un vaste corpus de récits de voyage en 4 langues, n'a été accessible que deux ans, suite au retrait du partenaire industriel alors impliqué. Sur Ortolang, on trouve plusieurs corpus parallèles téléchargeables (corpus GIEC, corpus Résolutions du Conseil de sécurité de l'ONU 1946-2015, ParCoGLiJ), mais aucun n'est consultable directement. Ainsi, outre le problème de pérennité des sites de projets, se pose aussi la question de la consultation des données pour un public de non spécialistes (ne maîtrisant pas nécessairement les formats tels que json, xml, tsv, tmx, etc.)

1 Projet mené par Andras Farkas, cf. https://sourceforge.net/projects/aligner/

2Accessible en ligne depuis WebAlignToolkit : http://phraseotext.univ-grenoble-alpes.fr/webAlignToolkit/

3 Développé par Michael Barlow : https://www.paraconc.com/index.html

Nous présentons ici des outils développés dans le cadre des projets ACR Grimm[4] et ParaTaxe[5], développés en *open source* et visant à combler cette lacune. Dans la section suivante nous présenterons l'interface Grimm Tradalign (grimm.unil.ch), développée par l'Université de Lausanne pour l'exploration de différentes traductions alignées d'un corpus de contes de Grimm. Nous présenterons ensuite une chaine de traitement élaborée à l'Université Grenoble Alpes pour le reformatage et l'alignement de ces corpus, ainsi qu'une interface en ligne nommée ParaTaxe Dashboard permettant d'importer des textes et de lancer cette chaine de traitement de manière automatisée. Nous conclurons sur les prochains développements concernant ces outils

L'interface Grimm Tradalign

L'interface Grimm Tradalign a été spécifiquement conçue pour permettre un affichage multiple. Contrairement à la plupart des concordanciers bilingues qui permettent d'afficher deux versions côte à côte, cette interface permet d'afficher en parallèle autant de versions que la taille de l'écran -- et le confort de lecture -- le permettent.

L'architecture est basée sur le moteur de la plate-forme TeiPublisher développée par la société Jinntec[6], dédiée à la publication de corpus encodés en XML TEI via le système de base de données XML eXist-db, lui-même publié en *open source*[7].

On désigne par "document" un ensemble de "versions", correspondant à un même texte original (p.ex. le conte *Hänsel und Gretel*). Chacune de ces versions est alignée avec un même texte pivot : pour les contes des Grimm, nous avons choisi l'édition de 1857. Les alignements sont représentés par des balises <*anchor*> dont l'attribut *corresp* pointe vers les identifiants de la version pivot.

```
</teiHeader>
<text>
  <body>
    <p>
      <anchor xml:id="KHM53_1857_gr_1"/>
      <s> Sneewittchen.</s>
    </p>
    <p>
      <anchor xml:id="KHM53_1857_gr_2"/>
      <s> Es war einmal mitten im Winter, und die Schneeflocken fielen wie Federr
      <anchor xml:id="KHM53_1857_gr_3"/>
      <s>Und wie sie so nähte und nach dem Schnee aufblickte, stach sie sich mit
      <anchor xml:id="KHM53_1857_gr_4"/>
      <s>Und weil das Rothe im weißen Schnee so schön aussah, dachte sie bei sich
      <anchor xml:id="KHM53_1857_gr_5"/>
      <s>„hätt ich ein Kind so weiß wie Schnee, so roth wie Blut, und so schwarz
      <anchor xml:id="KHM53_1857_gr_6"/>
      <s>Bald darauf bekam sie ein Töchterlein, das war so weiß wie Schnee, so rc
      <anchor xml:id="KHM53_1857_gr_7"/><s>Und wie das Kind geboren war, starb di
    </p>
```

```
</teiHeader>
<text>
  <body>
    <p>
      <anchor xml:id="KHM53_1846_ma_1" corresp="#KHM53_1857_gr_1"/>
      <s>Blancheneige.</s>
    </p>
    <p>
      <anchor xml:id="KHM53_1846_ma_2" corresp="#KHM53_1857_gr_2"/>
      <s>Un jour d'hiver, tandis que les flocons de neige tombaient du ciel comm
      <s>Et tandis qu'elle était ainsi occupée à coudre et à regarder tomber la
      <anchor xml:id="KHM53_1846_ma_4" corresp="#KHM53_1857_gr_4"/>
      <s>Et comme ce sang brillait si rouge dans la blanche neige, la reine se c
      <s>Si j'avais un enfant aussi blanc que cette neige, aussi rouge que ce sa
      <anchor xml:id="KHM53_1846_ma_6" corresp="#KHM53_1857_gr_6"/>
      <s>Peu de temps après, elle eut une petite fille blanche comme neige, rouç
      <anchor xml:id="KHM53_1846_ma_7" corresp="#KHM53_1857_gr_7"/>
      <s>Et dès que l'enfant fut né, la reine mourut.</s>
    </p>
    <p>
```

figure . 1 Représentation des alignement sous forme de balises <anchor> pointant vers la version pivot

Les segments sont représentés par des balises <s>. Une fois les textes encodés de cette manière, on peut les importer sur l'interface d'eXist-db, et les textes deviennent consultables en version multiparallèle dans l'interface de Tradalign, comme on le voit dans la figure 2. En cliquant sur un segment de n'importe quelle version, on voit alors s'afficher les segments correspondant en surbrillance. Les ajouts, par rapport à la version pivot, apparaissaient en bleu, et les omissions

4 Projet Grimm Tradalign (financement ACR 2023 – Alliance Campus Rhodanien, Projet ANR-15-IDEX-02)

5 Projet ANR - Idex UGA - IRGA2023 –ParaTAXE

6 cf. https://teipublisher.com/exist/apps/tei-publisher-home/index.html

7 cf. https://exist-db.org/exist/apps/homepage/index.html

sont marquées par des []. On peut aussi extraire des concordances en cherchant une expression dans le document.

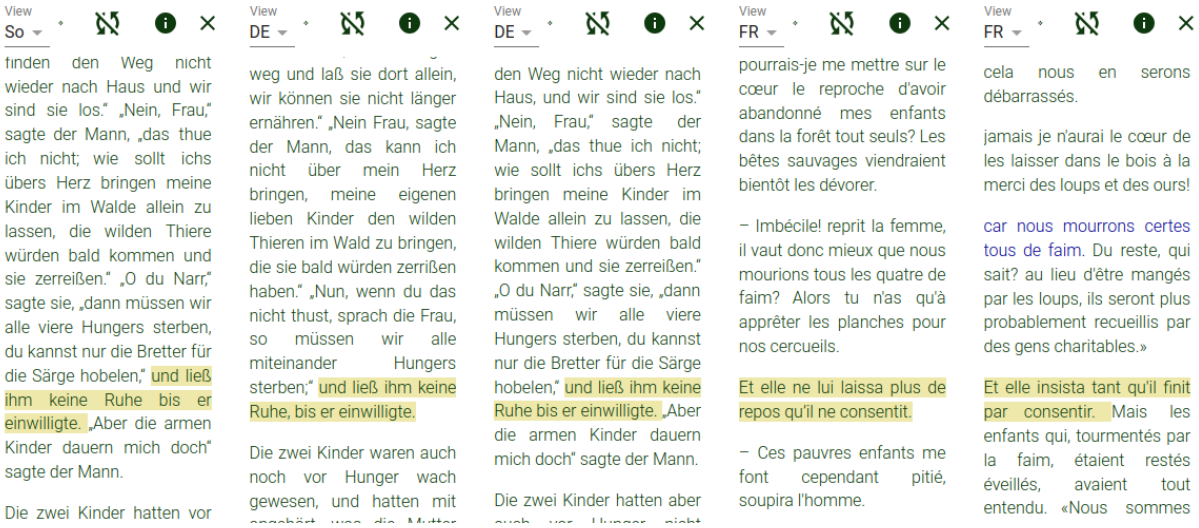


figure . 2 Figure 2Affichage multiparallèle des différentes versions d'un même document.

L'interface ParaTaxe Dashboard

La chaine de traitement développée par nos soins permet, à partir d'une liste de fichiers en format DOCX, TXT ou XML, accompagnés d'un tableau de métadonnées au format CSV, de générer les fichiers au format XML TEI correspondant, puis de lancer l'aligneur AIlign (Kraif, 2024) pour récupérer les fichiers avec les ancres appropriées. La compilation au format eXist-db et l'importation dans l'application sont également effectuées automatiquement. Via l'interface ParaTaxe, on peut ainsi gérer l'ajout et la suppression de documents (un ensemble de textes parallèles), afin de permettre une publication facilitée d'un ensemble de textes sur un serveur dédié, sans connaissance techniques spécifiques.

Perspectives

A terme, nous prévoyons d'enrichir l'outil pour permettre la publication de différentes collections sur un même serveur, et de le coupler à des outils d'extraction de statistiques lexicométriques (fréquences des lemmes et parties du discours, longueurs des phrases, taille du vocabulaire, fréquences des correspondances lexicales, segments répétés, etc.), utiles dans la perspective d'une traductologie de corpus telle que celle de Kraif & Roux (2022).

Références bibliographiques

Dekker R. H., Middell G. (2011). Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements. *Supporting Digital Humanities 2011*. University of Copenhagen, Denmark. 17-18 November 2011.

Dyer C., Chahuneau V., Smith N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

El-Bèze M., Richard C., Meyer R. (2006). Projet CARMEL : récits de voyages». in *Actes de TALN 2006*, Louvain-la-Neuve.

Gedzelman S., Zancarini J.-C. (2011). HyperMachiavel : un outil de comparaison de traductions. *Lingua e stile*, 2011, vol. XLVI, n° 2, pp. 247-266.

Kraif O. (2015). Multi-alignement vs bi-alignement : à plusieurs, c'est mieux ! , *Actes de TALN 2015, 22ème Conférence sur le Traitement Automatique des Langues Naturelles*, Caen, 22-25 juin 2015, pp. 255-266.

Kraif O., Roux P. (2022). Comparaison d'un texte original et de ses rétrotraductions : que disent les mesures textométriques ?, in Meng Ji Christine et Michael Oakes (Eds.), *Les nouvelles méthodologies de la traductologie de corpus : La révolution empirique en traductologie, Meta*, v. 67, no 1, Les Presses de l'Université de Montréal : Montréal.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas,* pages 135–144, Tiburon, USA, October. Springer.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.

Reboul M. (2022) Comparaison semi-automatique des traductions françaises de l'Odyssée d'Homère (1547-1955). In Didier Alexandre et Glenn Roe (éd.), *Cultures et pratiques savantes du numérique*, Classiques Garnier ⟨10.48611/isbn.978-2-406-12961-5⟩

Parler pour convaincre : routines discursives et stratégies de persuasion au marché *Fera ‘o Luni* de Catane

Chiara Manno
CENTRE D’ÉTUDES LINGUISTIQUES – CORPUS, DISCOURS ET SOCIÉTÉS (UR CEL)
UNIVERSITÉ JEAN MOULIN LYON 3
chiara.manno@univ-lyon3.fr

Introduction

Le marché de la *Fera ‘o Luni* [1] est un marché populaire très fréquenté qui se tient quotidiennement au centre-ville de Catane (Italie) et qui propose une grande variété de produits : alimentation, vêtements, articles domestiques entre autres. Il constitue un lieu d’interaction orale particulièrement riche, marqué par la pluralité des codes linguistiques et la mise en scène répétée de routines verbales.

Cette communication se propose d’analyser des séquences typiques du discours marchand issues de ce terrain, afin de dégager les formes linguistiques récurrentes et les modalités de persuasion ordinaires à l’œuvre.

Le travail s’inscrit dans la lignée des recherches sur les interactions en contexte commercial, qui ont mis en évidence l’existence de formes interactionnelles stabilisées - salutations, offres, énumérations, évaluations - tout en soulignant une adaptation au contexte local (Kerbrat-Orecchioni et Traverso, 2008 ; Traverso, 2001). Ces échanges, bien qu’asymétriques, sont co-construits, régis par des routines verbales et fortement ancrés dans leur environnement social et culturel. La dimension multimodale et prosodique de ces interactions a également été largement documentée.

En nous appuyant sur ces études, ainsi que sur les apports de la pragmatique interactionnelle (Kerbrat-Orecchioni, 2005), de l’analyse du cadre (Goffman, 1973) et de la persuasion ordinaire (Amossy, 2010), nous interrogeons les façons dont les vendeurs du marché mobilisent des routines expressives - vocatifs, répétitions, hyperboles - pour valoriser leurs produits, capter l’attention et engager l’interlocuteur dans un échange situé.

Corpus et méthodologie

Corpus

Les données analysées proviennent d’observations non participantes menées au marché de la *Fera ‘o Luni*. Elles incluent des interpellations adressées aux passants, des énumérations de produits, ainsi que des séquences de mise en valeur explicite de ces derniers. Les analyses ont

1 Littéralement : « la foire du lundi », aujourd’hui active tous les jours.

été faites à partir de transcriptions d'enregistrements vidéo, détruits après traitement, conformément aux principes éthiques et aux exigences du RGPD.

Méthodologie

L'analyse adoptée est qualitative et s'inscrit dans le cadre de la pragmatique interactionnelle. Les séquences sont étudiées à partir d'une grille inspirée des travaux de Goffman (1973), Traverso (2001) et Kerbrat-Orecchioni (2005), croisant les niveaux linguistiques, prosodiques et interactionnels. Les énoncés sont analysés selon leur fonction (ouverture, appel, valorisation), et leur inscription dans un répertoire socio-discursif propre aux vendeurs. Une attention particulière est portée aux formes d'insistance (énumérations, variation lexicale), aux modulations prosodiques (intonations montantes, cris scandés, ralentis expressifs), ainsi qu'à la façon dont ces marqueurs participent à une stratégie d'engagement.

Résultats

Les interactions observées au marché de la *Fera 'o Luni* révèlent une parole marchande à la fois codifiée, stylisée et stratégiquement déployée. Loin d'être de simples appels standardisés, les énoncés produits par les vendeurs constituent un répertoire interactionnel riche, ancré localement, mobilisé de manière répétée mais contextuellement ajustée. Ces formes discursives combinent des routines conversationnelles attendues à des stratégies de persuasion implicites et explicites, relevant à la fois de l'appel, de la valorisation du produit et de la co-construction du lien interlocutif.

Par exemple, les énoncés de type « *Ruoccoli, fagiolini, puseddaaa – talìa chi beddi* ! » (« brocolis, haricots, petits pois – regarde comme ils sont beaux ! ») combinent énumération rythmée, ancrage dialectal et qualification explicite. On y retrouve une structure d'appel classique en contexte marchand (Traverso, 2001), autour d'un schéma trinaire : mention du produit (*ruoccoli, fagiolini, pusedda*), adresse à l'interlocuteur (*talìa*), évaluation (*chi beddi*). Ces structures, proches des routines décrites dans les petits commerces de proximité (*ibid.*), sont ici amplifiées par la prosodie (voix haute, allongement vocalique) et la scénographie sonore du marché.

Certains vendeurs recourent à des verbes elliptiques, souvent en dialecte, comme dans *« Scalai »* (« j'ai baissé »), forme à la première personne qui, bien que non accompagnée d'un objet, active un scénario implicite de négociation anticipée. Le vendeur performe ici un rôle de générosité et d'effort commercial sans que l'acheteur n'ait eu à intervenir : il s'agit d'une stratégie persuasive fondée sur la mise en scène d'un geste favorable (Amossy, 2000).

La dimension performative est particulièrement visible dans les séquences où la parole se fait plus intense, presque musicale :

« Peperoni, patate, cipolla, melanzane, dai ! » (« Des poivrons, des pommes de terre, des oignons, des aubergines, allez ! ») : ce type d'énumération chantée, souvent scandée avec une prosodie montante puis relâchée sur *dai* (« allez »), vise à capter l'attention par le rythme et la musicalité de l'oralité, en associant contenu informatif et effet expressif.

Dans d'autres cas, le vendeur enchaîne des séquences longues, entre appel, information, répétition et relance :

« Prego signora, pregoo, pregoo… a n'euru i roccoli, segali e spinaci, talia chi ssu belli sti roccoli… le fragole bellissime ! » (« Allez-y madame, allez-y, allez-y… un euro les brocolis, les blettes et les épinards, regarde qu'est-ce qu'ils sont beaux les brocolis… les fraises trop

belles ! »). Ici, l'appel répété (prego, « allez-y », « je vous en prie ») s'alterne avec une annonce de prix, une énumération de légumes, une valorisation évaluative (fragole bellissime, « des fraises très belles »), et une interpellation visuelle (talia, « regarde »). Le discours prend la forme d'un flux contrôlé, alternant intensité et pauses, dans un jeu vocal proche d'une performance publique (Coupland, 2007).

Ce mode de mise en voix est également observable chez les vendeurs de vêtements, qui utilisent les marques comme preuve de qualité :

« D'a Pompea originali, tutti d'a Pompea... a ddu euro signora, taliamo ! » (« Des Pompea originaux, tous de la Pompea…deux euros madame, regardons ! »). La répétition de la marque (*Pompea*) fonctionne ici comme marqueur d'authenticité, afin de convaincre les potentiels acheteurs à conclure l'achat. Le verbe « regarder » à la forme exhortative « *taliamo* », invite les potentiels clients à s'assurer d'eux-mêmes de la qualité des produits.

Enfin, la sollicitation directe est fréquente :

« Signora, na cesta n'euro a voli ? » (« Madame, un panier un euro, vous le voulez ? ») ou « *Signora, i cavolfiori non li vuoleeee ? »* (« Madame, les choux-fleurs vous ne les voulez pas ? ») : ces énoncés engagent l'interlocutrice de manière frontale, tout en maintenant une symétrie contrôlée propre aux interactions de service (Kerbrat-Orecchioni, 2005 ; Traverso, 2001). L'insistance prosodique et la dislocation à droite du COD contribuent à la mise en relief du produit proposé. Le recours à l'italien, dans un terrain où le dialecte constitue la norme orale dominante, confère à l'énoncé une tonalité plus formelle, renforçant ainsi le caractère performatif de la proposition. Le client, bien que libre de refuser, est constamment remis dans la position d'un interlocuteur engagé, sollicité à travers des formes langagières qui relèvent autant de l'offre que de l'injonction déguisée.

Ces exemples illustrent un registre oral spécifique, où le discours marchand n'est ni chaotique ni aléatoire, mais fondé sur des schémas interactionnels reconnus et reconfigurés à chaque occurrence (Schank et Abelson, 1977). À l'instar des observations faites dans les commerces de proximité (Traverso, 2001), les routines du marché de Catane sont à la fois rituelles, stratégiques et expressives, et relèvent d'une compétence interactionnelle située, mobilisant le langage comme vecteur de lien autant que d'échange économique.

Perspectives

L'analyse invite à considérer les discours marchands comme des formes situées d'interaction ordinaire, à la fois ritualisées, expressives et stratégiques. En éclairant les usages langagiers du marché de Catane, elle contribue à la description de pratiques vernaculaires souvent marginalisées dans les études sur la parole publique.

Ce travail pourra être prolongé par une analyse comparative avec d'autres marchés, afin d'examiner les variations pragmatiques et prosodiques selon les contextes culturels, linguistiques et urbains. Il ouvre également sur des perspectives multimodales, en intégrant les gestes, postures et déplacements dans l'espace comme autant de ressources interactionnelles contribuant à l'efficacité du discours marchand.

Références bibliographiques

Alfonzetti, G. (2017). *Parlare italiano e dialetto in Sicilia*. Palerme, Centro di Studi Filologici e Linguistici Siciliani.

Amossy, R. (2010). *L'argumentation dans le discours*. Paris, Armand Colin (3ème édition).

Castiglione, M., Sardo, R. (2013). « Lingua, dialetto e scuola ». In G. Ruffino (dir.), *Lingue e culture in Sicilia*, 496-567. Palerme, Centro di Studi filologici e linguistici siciliani.

Coupland, N. (2007). *Style: Language variation and identity*. Cambridge, Cambridge University Press.

Doury, M. (2001). « Une discussion dans un commerce d'habitués ». *Les Carnets du Cediscor* [En ligne], 7, 1-14.

Drew, P., Heritage, J. (1992). *Talk at Work: Interaction in Institutional Settings*. Cambridge, Cambridge University Press.

Goffman, E. (1973). *La mise en scène de la vie quotidienne. Tome 1 : La présentation de soi*. Paris, Les Éditions de Minuit.

Kerbrat-Orecchioni, C. (2005). *Le discours en interaction*. Paris, Armand Colin.

Kerbrat-Orecchioni, C., Traverso, V. (2004). « Types d'interactions et genres de l'oral ». *Langages,* 153, 41-51.

Kerbrat-Orecchioni, C., Traverso, V. (dir.). (2008). *Les interactions en site commercial – Invariants et variations*. Lyon, Presses Universitaires de Lyon.

Schank, R. C., Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding - An inquiry Into Human Knowledge Structures*. Hove, Psychology Press.

Traverso, V. (2001). « Interactions ordinaires dans les petits commerces : éléments pour une comparaison interculturelle ». *Langage et société*, 95 (1), 5-31.

Pratiques de la métadonnée – le projet Open French Corpus

Céline Poudat[1], Mathilde Guernut[2], Marine Delaborde[3], Christophe Parisse[4],
[1] BCL, CNRS & Université Côte d'Azur
[2] STL, CNRS & Université de Lille
[3] LT2D, CY Cergy Paris Université
[4] Modyco, CNRS & Université Paris Nanterre
celine.poudat@univ-cotedazur.fr, mathilde23guernut@gmail.com, marine.delaborde@cyu.fr,
cparisse@parisnanterre.fr

Introduction

Le consortium CORLI travaille depuis maintenant quatre ans sur le projet *Open French Corpus* (Parisse et al. JLC 2023), qui vise à regrouper les corpus de langue française existants sous une forme standardisée, tant au niveau des formats que des métadonnées. L'initiative propose de centraliser ces corpus dans un espace commun accessible, accompagné d'outils spécifiques pour leur exploitation. Notre objectif est d'améliorer l'accès et l'utilisation de ces ressources, mais aussi leur cohérence, tout en garantissant la qualité des données à travers un processus de validation par la communauté scientifique. Nous inscrivons également ce travail en dialogue avec l'initiative européenne Eureco, qui partage la même ambition d'interopérabilité. Ce projet répond ainsi aux besoins croissants d'une utilisation harmonisée et la plus fiable possible des corpus dans divers domaines de recherche.

La présente communication se concentre spécifiquement sur la question des métadonnées. En effet, la mise en commun du contenu de corpus pour les rendre interopérables (interrogeables et téléchargeables dans un format commun) est une entreprise facilitée par les pratiques déjà anciennes de partages de données de linguistique. Par contre, la question de la mise en commun et de l'harmonisation des métadonnées se révèle d'une grande difficulté et complexité en raison de l'absence de normes et surtout de la variété des pratiques des concepteurs de corpus.

Complexité et politique en matière de métadonnées

Les métadonnées sont extrêmement utiles dans le cadre d'un projet comme celui du corpus ouvert de français (*Open French Corpus*). En effet, les futurs utilisateurs de l'OFC ne voudront pas nécessairement empiler en vrac des données de toutes origines comme cela peut l'être fait dans des très grands corpus issus du web, mais au contraire accéder à des domaines ou sous-domaines bien ciblés pour alimenter leurs travaux de recherche en linguistique.

Pour cela, il faudrait disposer d'un jeu de métadonnées idéal et le plus exhaustif possible, et déterminé de la même manière dans tous les corpus, genres et domaines. Un tel objectif est malheureusement peu réaliste. C'est le passage d'un jeu de métadonnées idéal à la réalité des pratiques qui nous occupe dans la présente proposition.

Notre point de départ est un jeu de métadonnées riche et exhaustif, réalisé en collaboration avec le consortium ARIANE (1.), et une évaluation des métadonnées des corpus déposés sur les plateformes de dépôt, à commencer par ORTOLANG (2.). Il s'agit alors pour nous de confronter l'idéal théorique d'un jeu de métadonnées standardisé et parfaitement structuré à la réalité du terrain, et à la complexité et l'hétérogénéité des métadonnées renseignées par les acteurs - quand elles le sont - tout en interrogeant la mise en œuvre d'un noyau élémentaire

acceptable et exploitable par les collègues. Nous conclurons notre propos en approfondissant notre réflexion, et en exposant les perspectives de notre travail.

Un jeu de métadonnées de corpus

Afin de développer un jeu de métadonnées adapté pour décrire les corpus et les textes, entendus au sens large, qu'ils rassemblent, nous avons d'abord constitué un groupe de travail chargé d'identifier les métadonnées pertinentes en fonction des usages discursifs (e.g. textes scientifiques, juridiques, journalistiques…). Comme nous le montrons dans la section qui suit, nous avons développé deux niveaux d'observation : le niveau du corpus (ou de la collection de documents) et le niveau du texte lui-même. En effet, un corpus peut être bien sûr hétérogène et contenir des textes de genres différents, tandis qu'il contiendra des informations spécifiques à sa constitution.

Parallèlement, depuis 2022, nous menons une collaboration fructueuse avec le consortium ARIANE en vue d'intégrer ces métadonnées au sein d'un thésaurus, en suivant l'exemple du thésaurus développé pour les textes littéraires (Galleron et al. 2021). Nous avons ainsi intégré les catégories du thésaurus d'Ariane qui nous semblaient pertinentes tout en ajoutant des catégories remontées par le groupe de travail d'une part, et représentées dans les corpus déposés sur Ortolang d'autre part. Pour l'heure, notre thésaurus s'articule suivant neuf catégories principales (https://opentheso.huma-num.fr/?idt=OFC-CORLI), chacune se déclinant en différentes sous-catégories spécifiques.

Conjointement à cette entreprise, nous avons procédé à une évaluation des corpus déposés sur Ortolang, de leurs formats d'une part et de leur possible intégration dans l'OFC, et des métadonnées posées sur les corpus et les textes, ce qui nous a permis d'observer les pratiques des chercheurs en la matière.

Evaluation des corpus déposés sur ORTOLANG

La plateforme ORTOLANG héberge des corpus hétérogènes, tant sur le plan thématique que formel, ce qui illustre la diversité des pratiques de recherche sur les données langagières. Pour le projet *Open French Corpus*, les formats exploitables sont les formats ouverts.

Petit panorama des corpus disponibles

À l'heure actuelle, parmi les corpus déposés sur ORTOLANG, la plupart (plus des deux tiers selon nos premières approximations) pourront intégrer l'OFC. Le format présent dans le plus grand nombre de fichiers (56,6%), mais aussi dans le plus grand nombre de répertoires (29.6%) est le format XML.

Si l'on observe les métadonnées des corpus déposés sur ORTOLANG, elles sont d'abord contraintes par le formulaire de dépôt, et doivent donc être décrites en mode et en genre(s). Ainsi, 56% des corpus sont oraux, 28% écrits et 16% multimodaux - avec certaines réserves puisque certains corpus sont multi-modes et ont choisi un mode de manière arbitraire.

ORTOLANG dispose d'un outil d'édition de métadonnées, mais il n'a pas été constitué de manière systématique et planifiée, mais au contraire de manière ad hoc en fonction des besoins exprimés par les déposants. Il fournit donc une base de travail intéressante mais qu'il faut revoir et faire évoluer. Par exemple, concernant les noms de genres, on observe différentes catégories génériques très générales dont le tableau 1 propose une approximation.

Genre	multimodal_corp	speech_corpora	written_corpora	Total général
journalistic_corpora	4	2	8	14
literature_corpora	2	5	14	21
natural_speech	18	34	2	54
new_medium	2	2	12	16
ordinary-professional_writing	1	1	15	17
scientific_corpora	3	3	16	22
scripted_speech	8	9	4	21
Total général	**38**	**56**	**71**	**165**

table 1. : Genre et mode dans les corpus d'ORTOLANG

Métadonnées des textes

Un corpus est généralement constitué d'un ensemble de textes potentiellement hétérogènes en genres, ce qui nécessite une attention particulière. C'est particulièrement le cas des grands corpus, qui peuvent avoir été constitués en vue d'observer un objet de recherche transversal, qui créera une divergence entre les métadonnées du corpus et celles des textes individuels. C'est ainsi, par exemple, qu'un corpus estampillé "juridique" peut contenir des romans, des textes philosophiques ou encore des articles encyclopédiques. L'objet de la recherche est sans aucun doute juridique, mais la catégorie "genre" est alors utilisée comme un label général du corpus.

De manière générale, le format de ces métadonnées ne suit aucune norme systématique et nécessite souvent un travail sur mesure. Même dans le cas où une norme est adoptée (par exemple utilisation d'un format TEI), une attention particulière doit être portée au contenu des métadonnées afin de les unifier, tout en tenant compte des usages discursifs qui développent leurs propres distinctions, comme le montre par exemple la figure 1.

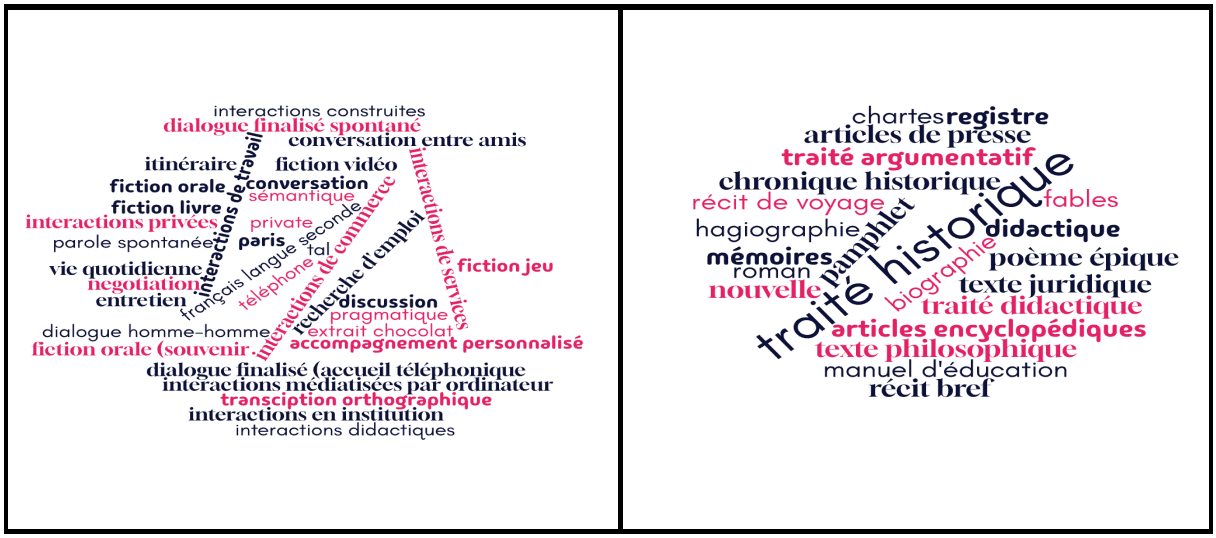


figure . 1 Nuage de mots des métadonnées distinctives des interactions orales et des textes juridiques

Conclusion et perspectives

Notre communication nous permettra de présenter l'état d'avancement de notre travail d'unification des métadonnées, ainsi que les noyaux minimaux que nous développons actuellement pour chaque usage discursif, en vue de leur exploitation dans un moteur de recherche à facettes.

Références bibliographiques

Galleron, I., Idmhand, F., Lavrentiev, A., Demonet, M.-L., & Réach-Ngô, A. (2021). *Décrire les textes dans le cadre d'une édition numérique*. https://shs.hal.science/halshs-03402679v1

Parisse C., Poudat, C., Badin, F., Benzitoun, C., Diwersy, S., et al.. CORLI : Un corpus ouvert du français -ou comment travailler à rassembler les briques existantes ?. Journées de Linguistique de Corpus, Jul 2023, Grenoble, France. ⟨hal-04255174⟩

Propositions pour une approche quantitative-qualitative des discours traitant des métiers dans l'*Encyclopédie* de Diderot et d'Alembert.

Alice Brenon[1], Denis Vigier[2]
[1] ICAR, ENS Lyon
[2] ICAR, Université Lyon 2
alice.brenon@ens-lyon.fr, denis.vigier@univ-lyon2.fr

Depuis les travaux fondateurs de J. Proust (entre autres 1962, 1965), les études consacrées à l'*Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*[1] (1751-1772) occupent une place de choix dans le champ la lexicographie historique française. Néanmoins, la part des études quantitatives s'y avère globalement faible (voir néanmoins Roe & al. 2016, Laramée 2017, Brenon & al. 2022 entre autres) et relativement tardive. De fait, la mise à disposition de l'œuvre entièrement numérisée et annotée, associée à une interface automatique de navigation et d'exploration au moyen de fonctionnalités statistiques, ne fut rendue possible qu'à la fin des années quatre-vingt grâce au programme ARTFL, puis en 2015 grâce au programme ENCCRE.

Qu'il s'agisse des études strictement qualitatives (conduites suivant des procédures d'analyse traditionnelles adossées à une lecture humaine de l'œuvre) ou des études quantitatives-qualitatives (mettant en jeu les plateformes évoquées *supra* ou d'autres plus élaborées : TXM, Perdido, Lexicoscope 2.0 ...), force est de constater que celles consacrées au traitement des métiers dans l'*Encyclopédie* sont peu nombreuses (voir néanmoins pour la première catégorie, e.a. B. Gille, 1952 ; J. Proust, 1957 ; P. Caye, 2009 ; Boussugue & Launay 2019). Situation d'autant plus étonnante si l'on considère que l'*Encyclopédie* fut le premier dictionnaire universel dans lequel « les métiers et les arts mécaniques se trouvent reliés au projet encyclopédique » (P. Caye, 2009 : 449-450).

Dans cette communication, on se propose d'approcher par des méthodes d'analyse textométrique le traitement des métiers dans l'*Encyclopédie*. Et cela en mettant en œuvre une double démarche : interne d'abord, puis externe-comparative ensuite.

Par approche *interne*, nous entendons une démarche contrastive entre sous-corpus constitués au sein des dix-sept volumes de texte numérisés et annotés de l'*Encyclopédie*. Ces sous-corpus ont été élaborés à partir de la métadonnée « macro-domaine » (voir *infra*) disponible dans notre version annotée de l'œuvre (Brenon & al. 2022). Notre analyse textométrique s'adossera pour l'essentiel aux calculs des spécificités et des cooccurrences dans TXM. Nous recourrons aussi à certaines fonctionnalités du Lexicoscope en vue de mettre en lumière, au moyen des arbres lexico-syntaxiques récurrents (ALR : voir entre autres Kraif & Diwersy 2014, Tutin & Kraif

1 éditée par Diderot, d'Alembert et Jaucourt (désormais l'Encyclopédie)

2016), certains motifs et séquences phraséologiques spécifiques au traitement des métiers dans l'*Encyclopédie*.

Par approche *externe-comparative*, nous entendons une étude contrastive conduite non au sein de l'*Encyclopédie* mais en comparant le traitement des métiers dans cette œuvre et celui proposé dans l'édition de 1743 du *Dictionnaire Universel François et Latin* de Trévoux (désormais DUFLT). Grand rival éditorial et idéologique de l'*Encyclopédie*, le DUFLT a connu, contrairement à sa rivale, six éditions[2] entre 1704 et 1771, chacune révisant et, le plus souvent, développant la précédente. L'édition sur laquelle nous nous appuierons est la seule dont nous disposons actuellement dans un format numérisé et annoté. Notre objectif consistera - en recourant aux mêmes plateformes et aux mêmes calculs textométriques que dans notre approche « interne » - à comparer le traitement discursif des métiers dans les deux œuvres. Cela afin de circonscrire les spécificités du discours tenu dans l'*Encyclopédie*, spécificités que nous mettrons en relation, dans notre dernière partie, avec l'objectif majeur que Diderot a poursuivi dans cette œuvre : participer à l'élaboration d'une « langue » et d'une « grammaire » des arts.

Les lignes qui suivent précisent et développent les grandes lignes du programme présenté *supra*.

Dans le cadre d'un programme de recherche financé par le LabEx ASLAN, nous avons accompli une classification automatique des domaines traités dans les articles de l'*Encyclopédie*. Nous nous sommes pour cela appuyés sur les plus de 60 000 articles de l'œuvre pourvus d'un « désignant », c'est-à-dire d'une information placée par les encyclopédistes immédiatement après la vedette de l'article et assignant à ce dernier un domaine (*médecine, horlogerie …*), voire plusieurs. Cependant, l'étude de la distribution des domaines dans l'*Encyclopédie* tels que formulés par les encyclopédistes, montre qu'un très grand nombre d'entre eux est associé à un nombre très faible d'articles, voire souvent à un seul. Cela nous a donc amené à regrouper ces derniers en quarante-quatre « macro-domaines » afin de rendre possible l'entraînement de notre modèle de classification. Nous avons réalisé cette tâche en nous appuyant sur les regroupements opérés par ENCCRE. C'est cette annotation que nous mobiliserons pour nos études textométriques.

Dans la première partie de notre communication, après avoir précisé comment les encyclopédistes définissent et articulent les concepts de science, d'art et de métier dans leur dictionnaire, nous mettrons en œuvre une étude textométrique des discours traitant des métiers dans cette œuvre à partir de plusieurs sous-corpus de travail constitués pour cette tâche. Nous contrasterons ainsi successivement les sous-corpus d'articles traitant des macro-domaines suivants : *métiers de l'alimentation, métiers du bois, métiers du cuir et des peaux, métiers du métal (…), métiers du papier, métiers du tissu et de l'habit* avec l'ensemble des autres macro-domaines ne relevant pas des arts mécaniques (*Anatomie, Beaux-Arts, Belles-Lettres etc.*). Les fonctionnalités de TXM auxquelles nous recourrons pour une telle analyse seront le calcul des spécificités et le calcul des cooccurrences, en vue d'explorer et de cerner le profil combinatoire (Blumenthal, 2008) des termes les plus spécifiques à ces métiers. Nous prolongerons notre étude en recourant aux arbres lexico-syntaxiques récurrents que permettent d'élaborer la plateforme Lexicoscope 2.0. Notre objectif sera de mettre au jour des motifs spécifiques susceptibles d'être reliés au traitement des métiers, par D. Diderot en particulier, dans notre troisième partie.

2 Si l'on excepte les deux éditions dites « nancéiennes » (1734, 1738-42). Les éditions ici considérées sont 1704, 1721, 1732, 1743, 1752, 1771.

On peut ici signaler que nos premières études exploratoires font d'ores et déjà apparaître, parmi les termes à forte spécificité, outre ceux (attendus) qui relèvent des matières travaillées dans les métiers considérés, certains termes abstraits tels que *sorte*, *manière* qui signalent la prévalence d'opérations de catégorisation conceptuelle mettant en jeu des *enclosures* ou *hedges* (Kleiber & Riegel, 1978), traitables sous le chef de la sémantique du prototype (Kleiber, 1990). Voici deux exemples :

CHERCHE-FICHE, (*Serrur.*) c'est **une sorte de** pointe acérée dont la tête forme un tour d'équerre, & est ronde de même que le reste du corps de cet outil. (…)

BATTRE L'OR, L'ARGENT, LE CUIVRE, (Art de battre l'or) (…) La partie *M N O P* peut contenir du feu. C'est **une espece de** vaisseau de fer.

Le second exemple intègre des lettres (*M, N, O, P*) qui désignent autant de régions sur une planche correspondante, ce qui nous conduira à approfondir le rapport spécifique qu'entretiennent les termes de métiers avec les onze volumes de planches.

Dans notre deuxième partie, nous procéderons à une étude contrastive entre les métiers regroupés par sous-classes dans l'*Encyclopédie* (cf. *supra*) et ceux que nous aurons classifiés dans l'édition de 1743 du DUFLT. Après avoir cursivement expliqué comment nous aurons procédé pour une telle classification, nous reprendrons les étapes suivies dans notre partie 2 : calcul des spécificités prolongés par des calculs de cooccurrence, mise en œuvre des ALR. Notre objectif majeur consistera à déterminer dans quelle mesure nos résultats nous permettent d'une part de décider si les spécificités et les motifs phraséologiques mis au jour dans notre première partie sont spécifiques à l'*Encyclopédie* ou bien relèvent d'un traitement plus vaste des métiers dans les dictionnaires universels du XVIIIe s. D'autre part, si nos résultats nous permettent d'esquisser des pistes pour une réflexion quant aux différences d'approche du traitement des métiers entre ces deux œuvres majeures de ce même siècle.

Enfin, notre troisième partie se fixe pour objectif de mettre en relation les résultats de nos approches qualitatives-quantitatives avec les objectifs qu'a poursuivis Diderot dans la rédaction des articles traitant des métiers (et plus largement des arts mécaniques). En effet, un des obstacles majeurs auquel s'est heurté cet auteur dans l'*Encyclopédie* est la relative rareté d'une documentation technique de qualité traitant des arts mécaniques et des métiers. Il lui a donc fallu se familiariser avec le lexique propre à chacun des arts dont il rendait compte. Il lui a aussi fallu se colleter avec l'imperfection de la langue des arts mécaniques. Il développe ce point notamment dans l'article ART : « J'ai trouvé la langue des *Arts* très-imparfaite par deux causes ; la disette des mots propres, & l'abondance des synonymes. Il y a des outils qui ont plusieurs noms différens; d'autres n'ont au contraire que le nom générique, *engin, machine*, sans aucune addition qui les spécifie : quelquefois la moindre petite différence suffit aux Artistes pour abandonner le nom générique & inventer des noms particuliers ; d'autres fois, un outil singulier par sa forme & son usage, ou n'a point de nom, ou porte le nom d'un autre outil avec lequel il n'a rien de commun. » (ART, vol. 1, p. 716). Autrement dit, Diderot fait face à une multitude de dialectes[3] techniques qui coexistent les uns avec les autres. L'un des objectifs qu'il s'est fixé dans son entreprise intellectuelle vis-à-vis des arts et des métiers a donc consisté à élaborer un langage et une grammaire qui rendent possible l'émergence d'un discours adéquat. Il y a là un

3 Ce que note Diderot va exactement dans ce sens « dans la langue des Arts, un marteau, une tenaille, une auge, une pelle, &c. ont presque autant de dénominations qu'il y a d'Arts. La langue change en grande partie d'une manufacture à une autre. »

enjeu intellectuel et humain de premier plan qui se situe au cœur du projet philosophique et politique de l'*Encyclopédie*, et que nous essaierons d'éclairer par cette étude exploratoire.

Bibliographie citée :

Bertrand, G. (1952), « L'Encyclopédie, dictionnaire technique », *Revue d'histoire des sciences et de leurs applications*, 5 (1), 26-53;

Blumenthal, P. (2008), « Combinatoire des prépositions. Approche quantitative », *Langue Française,* 157, 37-51.

Boussuge, E. & Launay, F. (2019) « La description des arts dans l'*Encyclopédie* », *Recherches sur Diderot et sur l'Encyclopédie*, 54, 181-253.

Brenon, A., Moncla, L., Mcdonough, K. (2022), *Classifying encyclopedia articles: Comparingmachine and deep learning methods and exploring their predictions. Data and Knowledge Engineering*, 142, pp.102098.

Caye, P. (2009), « Le chantier sans maître. L'Encyclopédie et la question de la technique », *Dix-huitième siècle,* n° 41(1), 449-467Kleiber, G. (1990), *La sémantique du prototype. Catégories et sens lexical*, PUF : Paris

Kleiber, G. (1990), *La sémantique du prototype. Catégories et sens lexical*, PUF : Paris

Kleiber, G. & Riegel, M. (1978), « *Les « grammaires floues* » », in R. Martin (éd.), *La notion de recevabilité en linguistique*, Paris : Klincksieck, 67-123.

Kraif, O, Diwersy, S. (2014). Exploration de profils combinatoires à l'aide de lexicogrammes sur un corpus analysé : A Case Study in the Lexical Field of Emotions, in *Emotions in Discourse*. Berlin : Peter.

Laramée, F.-D, (2017). La production de l'espace dans l'Encyclopédie Portraits d'une géographie imaginée. *Document numérique,* . 20(2), 159-177

Proust, J. (1957), « La documentation technique de Diderot dans l'*Encyclopédie* », Revue d'Histoire littéraire de la France, 3, 335-352

Proust, J. (1962), *Diderot et l'Encyclopédie*, Paris, Armand Colin.

Proust, J. (1965), *l'Encyclopédie*, Paris, Armand Colin.

Roe, G., Gladstone, C., Morrissey, R. (2016), *Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie*, in *Frontiers in Digital Humanities* 2.

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

Quand le mouvement naît de la stativité : étude expérimentale du mouvement fictif en chinois standard

Jiayi LI [1]
[1] Laboratoire Lattice, Université Sorbonne Nouvelle
lijiayi25@hotmail.com

Introduction

Le mouvement fictif constitue un sujet de débat vif en linguistique cognitive depuis les années 1970. Ce phénomène se définit comme la description dynamique d'une scène spatiale statique, souvent réalisée au moyen de verbes de mouvement. Par exemple, dans *The fence descends from the plateau to the valley* (Talmy, 1983), la figure inanimée *fence* « clôture » ne se déplace pas réellement, mais elle est décrite par le verbe de mouvement *descend* « descendre ».

Les recherches actuelles sur le mouvement fictif se concentrent principalement sur trois axes : théorique, empirique et comparatif. Plusieurs travaux ont exploré ce phénomène d'un point de vue théorique, sous diverses dénominations, notamment les travaux de Talmy (1983, 1996), Langacker (1986), ainsi que Matsumoto (1996). Des approches expérimentales, monolingues ou multilingues, ont également été menées par Rojo et Valenzuela (2003), Matlock (2006), Blomberg (2015), Stosic et al. (2015), entre autres.

La présente étude porte sur l'expression du mouvement fictif en chinois standard, sur la base de données expérimentales. Celles-ci ont été recueillies au moyen d'une tâche d'élicitation fondée sur des images, afin d'inciter les participants à produire des énoncés contenant du mouvement fictif. Elle a pour l'objectif d'analyser l'usage oral de ce phénomène à travers des productions spontanées en chinois standard. Des résultats comparables ont été obtenus en français, suédois, thaï, italien, allemand et serbe (Blomberg, 2015 ; Stosic et al., 2015), selon une méthodologie unifiée. Pour le chinois standard, Gong et Zheng (2018) ont mené une première expérimentation à petite échelle (10 locuteurs natifs) à partir du même protocole, dans le but d'identifier les conditions les plus propices à la production du mouvement fictif. Notre étude vise à approfondir et à consolider ces résultats en nous appuyant sur un échantillon plus large et plus diversifié (40 locuteurs natifs), tout en intégrant le chinois à une perspective comparative plus vaste. Elle contribue ainsi à une meilleure compréhension des spécificités linguistiques du chinois standard dans le cadre du mouvement fictif interlinguistique.

Corpus et méthodologie

Corpus

Le corpus utilisé dans cette étude provient d'une expérimentation d'élicitation conçue par Jordan Zlatev et Johan Blomberg de l'université de Lund. Cette expérimentation se compose de trente-huit images, dont deux images d'entrainement, douze images de distracteurs et vingt-quatre images cibles. Ces dernières contiennent toutes un objet linéaire et étendu dans l'espace, positionné par rapport à un point de repère ou au fond (arrière-plan). Les images cibles ont été

créées en fonction de deux groupes de variables : l'accessibilité de la figure[1] pour l'humain et le point de vue adopté. Voici quatre images cibles extraites de l'expérimentation. Sur l'axe horizontal, nous pouvons voir que les figures 1 et 2 ainsi que les figures 3 et 4 contiennent les mêmes figures, à savoir « chemin » et « clôtures ». Le « chemin » permet le mouvement humain alors que la « clôture » ne le permet pas. Sur l'axe vertical, les figures 1 et 3 sont conçus du point de vue de la première personne (égocentré), tandis que les figures 2 et 4 adoptent celui de la troisième personne (allocentré) :

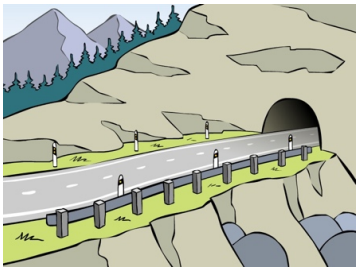


figure . 1 Figure 3Accessible/Égocentré

figure . 2 Figure 4Accessible/Allocentré

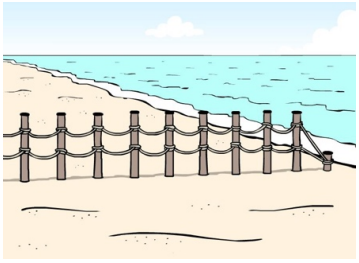


figure . 3 Figure 5Non accessible//Égocentré

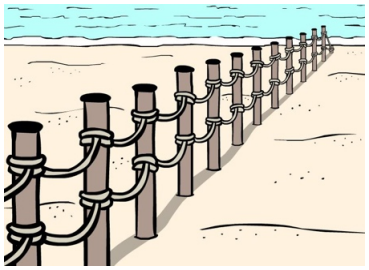


figure . 4 Figure 6Non accessible/Allocentré

1 La figure est l'objet saillant, mobile ou stationnaire, dans un évènement de mouvement. (Talmy, 1985)

L'objectif principal de cette expérimentation est d'étudier le mouvement fictif dans différentes langues, en mettant l'accent sur deux variables : l'accessibilité de la figure pour l'humain (accessible vs non accessible) et le point de vue adopté (égocentré vs allocentré), susceptibles d'influencer la production du mouvement fictif. L'étude vise également à examiner si l'importance relative de ces facteurs varie d'une langue à l'autre.

Méthodologie

Nous avons finalement mené l'expérimentation du côté de la langue chinoise auprès de 40 locuteurs natifs, dont 20 hommes et 20 femmes, âgés de 20 à 34 ans au moment de l'expérimentation. Les participants, originaires de différentes régions de Chine, travaillent ou étudient dans divers domaines. Aucun des participants n'avait de connaissances préalables sur le mouvement fictif avant l'expérimentation, y compris ceux issus du domaine de la linguistique. Six participants ont réalisé l'expérimentation en présentiel dont cinq à Paris et une à Tianjin (Chine), tandis que les autres ont participé en visioconférence via Zoom.

Toutes les séances ont été enregistrées sous forme de vidéo. Nous avons ensuite effectué une transcription littérale, en conservant l'intégralité du contenu sans aucune modification.

Le traitement des données a été réalisé à l'aide d'un tableau Excel. Nous y avons saisi les informations suivantes : numéro et nom de participant, description brève et type d'image (entraînement/distracteur/cible), description de l'image par le/la participant(e), figure impliquée dans l'image, type de figure, point de vue, type de mouvement impliqué (réel, fictif ou sans mouvement), et marqueur de mouvement utilisé.

Une fois le codage terminé, nous avons procédé à l'analyse en nous concentrant sur les questions suivantes :

- Sous quelle combinaison de conditions (figure accessible ou non accessible pour l'humain, point de vue à la 1ère personne ou à la 3ème personne), le mouvement fictif se produit-il le plus fréquemment par les participants chinois ? Les résultats de Gong et Zheng (2018) suggèrent une préférence marquée pour les images conçues du point de vue de la 1ère personne et impliquant une figure permettant le mouvement humain (par exemple, route, pont). Cette tendance se confirme-t-elle dans notre corpus plus étendu et diversifié ?
- Quels sont les marqueurs de mouvement les plus fréquents ?
- Si les participants chinois ne produisent pas de mouvement fictif, quels moyens linguistiques ont-ils emploient pour décrire les images cibles ?
- Dans quelle mesure les données issues des participants chinois reflètent-elles ou s'écartent-elles des tendances observées dans les autres langues étudiées par Blomberg & Zlatev (2014) et Stosic et al. (2015) ?

Notre première hypothèse est que les participants chinois, à l'instar de ceux des autres langues, produisent davantage d'expressions du mouvement fictif dans certaines configurations spécifiques. Nous supposons également que le nombre d'expressions du mouvement fictif produites par les locuteurs chinois est inférieur à celui observé chez les participants français, comme le suggèrent notre lecture sur les travaux précédents dans le domaine. Nous espérons en particulier identifier les moyens linguistiques, autres que les marqueurs de mouvement, que les participants chinois mobilisent pour décrire les images cibles.

Résultats

Les résultats obtenus indiquent que le mouvement fictif est relativement rare en chinois comparé aux autres langues étudiées. Une explication possible, mise en évidence par l'approche expérimentale, est que le chinois dispose d'autres moyens linguistiques pour décrire les

relations spatiales de manière statique, sans recours au mouvement fictif. Par exemple, face à une même image représentant des poteaux électriques alignés vers l'horizon, certains participants ont utilisé le verbe 蔓延 *mànyán* « s'étendre » (exemple 1), tandis que d'autres ont opté le classificateur 排 *pái* « rangée » (exemple 2), indiquant également une relation spatiale. Cette observation suggère une concurrence entre différentes structures syntaxiques, et montre que l'usage du mouvement fictif est partiellement contraint par les ressources lexicales et grammaticales propres à chaque langue :

1. 电线杆是由西向东地往前**蔓延**。

diànxiàngān	*shì*	*yóu-xī-xiàng-dōng*	*de*	*wǎng-qián*	***mànyán***
poteau électrique	être	depuis-ouest-vers-est	SUB	vers-devant	**s'étendre**

« Les poteaux électriques **s'étendent** d'ouest en est. »

2. 路中间有一**排**电线杆。

lù=zhōngjiān	*yǒu*	*yì-**pái***	*diànxiàngān*
route=milieu	avoir	un-**CL**	poteau électrique

« Au milieu de la route se trouve une **rangée** de poteaux électriques. »

Liste d'abréviation

CL	classificateur
SUB	subordinateur

Références bibliographiques

Blomberg, J., & Zlatev, J. (2014). Actual and non-actual motion: Why experientialist semantics needs phenomenology (and vice versa). *Phenomenology and the cognitive sciences*, *13*, 395-418.

Blomberg, J. (2015). The expression of non-actual motion in Swedish, French and Thai. *Cognitive Linguistics*, *26*(4), 657-696.

Gong, S. P., & Zheng, X. Y. (2018). Producing Fictive Motion Sentences in a Picture-Elicitation Task: A Pilot Study. *International Journal of Knowledge and Language Processing*. *9*(2) : 20-32

Langacker, R. W. (1986). Abstract motion. in Nikiforidou V., Van Clay M., Niepokuj M. et Feder D., *Proceedings of the 12th Annual Meeting of the Berkeley Linguistics Society. Berkeley Linguistics Society*, pp. 455-471.

Matsumoto, Y. (1996). Subjective motion and English and Japanese verbs. *Cognitive Linguistics*, *7*(2) : 183-226.

Matlock, T. (2006). Depicting fictive motion in drawings. *Cognitive Linguistics: Inverstigations across languages, fields, and philosophical boundaries*. Amsterdam: John Benjamins, pp. 67-85.

Rojo, A., & Valenzuela, J. (2003). Fictive Motion in English and Spanish. *International Journal of English Studies*, *3*(2), 123-150.

Stosic, D., Fagard, B., Sarda, L., & Colin, C. (2015). Does the road go up the mountain? Fictive motion between linguistic conventions and cognitive motivations. *Cognitive processing*, *16*, 221-225.

Talmy, L. (1983). How language structures space. In *Spatial orientation: Theory, research, and application*. Boston, MA: Springer US, pp.225-282.

– (1996). Fictive motion in language and “ception”. in Bloom P., Peterson M.A., Nadel L., et Garrett M.F. (Eds.), *Language and space*. Cambridge MA: MIT Press, pp. 211-276.

Quantifier le genre dans les références aux animaux en anglais

Martin Rioult[1,2]
[1] Laboratoire LIDILEM, Université Grenoble Alpes
[2] Laboratoire CLIMAS, Université Bordeaux Montaigne
rioultma@univ-grenoble-alpes.fr

Introduction

Un cafard est toujours un « il » en français, mais qu'en est-il en anglais ? Est-il un « he », un « she » ou un « it » ? En français, le genre grammatical des noms est une propriété fixe qui suit des règles largement arbitraires d'un point de vue du sens. En anglais, au contraire, le genre est dit « sémantique ». Pour les humains, il correspond généralement au sexe biologique ou à l'identité de genre des référents, tandis que pour les inanimés, le genre par défaut est le neutre. Pour les animaux, en revanche, il existe un « haut degré de variabilité » dans l'assignation du genre (Corbett, 1991 : 12). Il est ainsi possible, et même fréquent, d'associer le masculin, le féminin ou le neutre pour un même nom d'animal.

Le genre en anglais est un système de classification hiérarchique des entités qui repose sur une vision anthropocentrée du monde (Cotte, 1999 ; Gardelle, 2006 : 22). Cette hiérarchie place les humains à son sommet, typiquement repris par le masculin ou le féminin, et les inanimés à sa base, repris par le neutre. Les animaux y occupent une place ambivalente et instable car aucun genre ne leur correspond d'office. De surcroît, ils ne constituent en rien une catégorie homogène. Par défaut, it est plus fréquent dans les références aux animaux car ils sont non humains (Gardelle, 2023 : 398). he et she peuvent être utilisés si le locuteur connait le sexe biologique du référent et souhaite encoder cette information en discours (ex. 1). Le neutre n'est toutefois pas exclu dans ce cas (Gardelle, 2012 : 21). Le critère du sexe est cependant loin d'expliquer toutes les occurrences du masculin et du féminin. Un autre facteur intervient : le lien affectif ressenti par le locuteur envers le référent (ex. 2). Dans ce cas, le masculin (et beaucoup plus rarement le féminin) est utilisé afin de marquer cette proximité. Ce critère est décrit comme déterminant dans l'attribution du genre par de nombreux linguistes (Joly, 1975 ; Mathiot & Roberts, 1979 ; Gardelle, 2006 ; Morris, 2021).

1. The [female cockroaches] all rejected this male's mating attempts, they knew **he** was ill I guess! (cockroach-forum-Roach Forum-130)
2. 'Gave **him** the Aladdin treatment': Why kill a cockroach when you can give **it** a first-class flight instead? (cockroach-news-We Got This Covered-49)

Si le fonctionnement général du genre est aujourd'hui bien établi (Joly, 1975 ; Gardelle, 2006 ; Morris, 2021), les tentatives de quantification du phénomène dans les références aux animaux restent rares (MacKay & Konishi, 1980 ; Gardelle, 2012) et ne permettent pas d'établir plus finement les tendances d'usage. C'est donc le but de nos recherches actuelles.

Cette communication propose d'explorer plus spécifiquement comment mettre en œuvre une telle quantification grâce à une étude de corpus. Nous nous pencherons dans un premier temps sur le type de corpus pouvant être utilisé pour mener à bien notre recherche et déterminerons que les corpus existants sont trop lacunaires. Nous nous demanderons ensuite comment

constituer un corpus ad-hoc et examinerons les difficultés qu'une telle entreprise pose. Enfin nous discuterons de la méthode employée pour analyser ce corpus.

Corpus et méthodologie

Corpus : nécessité d'un corpus ad hoc

Collecter des données dans un des grands corpus de référence de l'anglais est inadéquat pour étudier le genre de manière quantitative et selon plusieurs variables (cf. Méthodologie infra). C'est ce qu'a montré une étude pilote que nous avons menée dans le COCA[1] — corpus de référence de plus d'un milliard de tokens, qui se veut représentatif de l'anglais américain contemporain. Ce corpus s'est révélé insuffisant car les références pronominales à certains animaux de l'étude étaient trop rares : on n'en comptait qu'environ cinquante pour le cafard, par exemple. Ainsi, les variables prévues n'ont pas pu être testées faute de chiffres suffisants et certaines données provenaient trop souvent des mêmes sources. Cela pose problème dans la représentativité des données : certains locuteurs et certaines locutrices ont plus de poids que d'autres dans le corpus. Les conclusions qui ont pu être tirées des données extraites du COCA sont donc apparues comme trop fragiles. D'autres corpus de grande taille comme le English Web Corpus (enTenTen) ont également été consultés mais ils ont tous présenté des lacunes : ils ne permettent souvent pas d'effectuer un tri par genre discursif et ne donnent accès qu'à un contexte restreint.

Partant de ce constat, l'élaboration d'un corpus ad hoc est la seule solution adéquate. Celui-ci est thématique : il ne contient que des textes contenant les noms d'animaux ciblés (ex : *cockroach*). Cela permet d'obtenir de grands jeux de données pour chaque animal retenu dans l'étude et ainsi de rendre possible l'analyse des différentes variables pouvant influer sur la sélection du genre.

Méthodologie

Afin de faciliter et d'harmoniser le traitement des données, les textes constituant le corpus sont collectés sur Internet sur la période 1990-2025 grâce à différents agrégateurs (*Europresse* pour la presse et *Google Scholar* pour les articles académiques) et divers forums en ligne via une recherche par mot-clé du nom de l'animal dans le titre. Selon les sources, ces textes sont obtenus manuellement ou automatiquement via des scripts en Python.

L'étude cible spécifiquement l'usage du genre via les pronoms personnels de troisième personne du singulier de l'anglais : he, she, it et leurs dérivés. Ces pronoms sont extraits des jeux de données grâce au concordancier AntConc (Anthony, 2024). Chaque concordance est ensuite transférée dans un tableur Excel pour en permettre le tri et l'annotation manuels. Les pronoms personnels apparaissant nécessairement en grand nombre dans tout texte, un tri initial est effectué : seules les occurrences faisant référence à l'animal cible sont conservées. En moyenne, 80 à 90% des occurrences sont systématiquement éliminées du jeu de données. A ce stade, le corpus compte 1 203 textes contenant 9 484 pronoms dont 1 072 renvoient à un animal cible (cf. Table 1).

1 Corpus of Contemporary American English (Davies, 2008).

	Ecrits académiques	Presse écrite	Forums en ligne	Total
Textes	20	581	602	1 203
Tokens	148 211	442 307	183 851	774 369
Pronoms bruts	837	5 000	3 647	9 484
Pronoms triés	116	204	752	1 072

table 1. : Taille actuelle du corpus pour un seul animal : le cafard.

Chaque occurrence est ensuite annotée selon une dizaine de variables susceptibles d'influencer le genre utilisé. Ces variables sont :

- Propres au référent : sexe biologique, nom, relation à l'homme.
- Propres à la syntaxe de l'énoncé : référence générique ou spécifique ; antécédent : nature, fonction, rôle sémantique ; pronom anaphorique : fonction, rôle sémantique ; position du pronom anaphorique par rapport à l'antécédent.
- Propres à la relation entre le locuteur et le référent : attitude du locuteur envers le référent.

On étudie ensuite les interactions significatives entre le genre utilisé et chaque variable grâce au test de χ^2. Les résidus de Pearson sont calculés afin d'identifier les données contribuant le plus à la valeur de χ^2 obtenue. Une étude qualitative vient compléter l'approche quantitative afin de mieux comprendre les interactions entre les différentes variables.

Enjeux et limites d'un corpus ad hoc sur le genre en anglais

Le corpus ad-hoc, de taille plus modeste que les grands corpus de référence, est constitué de différents genres discursifs, incluant à ce stade : écrits académiques, presse écrite et forums en ligne. Ces genres ont été sélectionnés pour présenter une diversité de points de vue sur les animaux. Par exemple, il est attendu que les écrits académiques privilégient majoritairement it, traduisant une certaine neutralité affective. Les écrits légaux ou administratifs, trop proches des textes académiques sur le plan du genre grammatical, ont été écartés. Les genres retenus doivent par ailleurs être accessibles au préalable en format informatisé.

Le nombre d'espèces retenues dans l'étude est nécessairement restreint, d'une part pour des raisons de faisabilité et d'autre part afin d'assurer un volume de données maximal pour chacune. De surcroît, les animaux ont été sélectionnés selon différents critères à même de modifier la sélection du genre grammatical. Par exemple, ils occupent différentes places sur un supposé axe animaux supérieurs / animaux inférieurs (Quirk et al., 1985 : 317; Siemund, 2008 : 164). La sélection vise aussi à refléter la diversité des espèces animales (mammifères, oiseaux, poissons, insectes, etc.).

Premiers résultats

Le corpus ainsi constitué permet d'obtenir des résultats nouveaux, robustes par leur appui quantitatif. A ce stade de l'étude, ils suggèrent une forte interdépendance entre genre discursif et genre grammatical : le neutre domine très largement dans les écrits académiques et la presse alors qu'il est bien moins fréquent dans les forums (cf. Figure 1). Ce résultat n'est pas inattendu car la sélection du genre dépend en grande partie de l'attitude du locuteur envers le référent. En effet, si le locuteur est plus détaché dans les écrits académiques ou la presse, il est potentiellement beaucoup plus investi dans les forums. Ces résultats seront complétés par des études qualitatives afin de les comprendre plus finement.

figure . 1 Interactions entre genre grammatical et genre discursif dans le corpus « cockroach ». Une interaction significative est de couleur bleue ou rouge.

Références bibliographiques

Anthony, L. (2024). *AntConc* (Version 4.3.1) [Logiciel]. Waseda University. https://www.laurenceanthony.net/software/AntConc

Corbett, G. G. (1991). *Gender*. Cambridge University Press.

Cotte, P. (1999). Le genre est une métalangue. Dans *Féminin/masculin: Littératures et cultures anglosaxonnes* (pp. 65–75).

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca/

Gardelle, L. (2006). *Le genre en anglais moderne (XVIe siècle à nos jours): Le système des pronoms*. Thèse de doctorat non publiée.

Gardelle, L. (2012). Pronominal gender in references to animals: A statistical analysis of the influence of six variables in discourse. *Anglophonia/Sigma*, *16 (32)*, 11–23. https://doi.org/10.4000/anglophonia.127

Gardelle, L. (2023). Pronoun activism and the power of animacy. Dans *The Routledge Handbook of Pronouns* (pp. 394–408). Routledge.

Joly, A. (1975). *Toward a Theory of Gender in Modern English*. Presses Universitaires de Lille.

MacKay, D. G., & Konishi, T. (1980). *Personification and the pronoun problem*. *2–3*(3), 149–163.

Mathiot, M., & Roberts, M. (1979). Sex Roles as Revealed Through Referential Gender in American English. Dans *Ethnolinguistics: Boas, Sapir and Whorf Revisited* (pp. 1–48). Mouton Publishers.

Morris, L. (2021). *Gender in Modern English: The System and Its Uses*. Les Presses de l'Université Laval. (Publication originale : 1991)

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (Eds.). (1985). *A Comprehensive Grammar of the English Language*. Longman.

Siemund, P. (2008). *Pronominal Gender in English: A Study of English Varieties from a Cross-Linguistic Perspective* (1st ed.). Routledge. https://doi.org/10.4324/9780203455944

Quelle interaction par fichier interposé ?

Marie-Paule Jacques [1]
[1] CLLE, Université Toulouse Jean-Jaurès
marie-paule.jacques@univ-tlse2.fr

Introduction

Notre proposition concerne les « échanges » entre un-e étudiant-e et son encadrant pour un travail universitaire, qui circulent sous forme d'un fichier, au sein duquel l'encadrant-e – et plus rarement l'étudiant-e, comme on le verra – inscrit des commentaires à destination de l'autre partie.

Ces commentaires ont été étudiés dans le cadre scolaire, où ils ont reçu diverses appellations : annotations, commentaires, interventions enseignantes (par ex. Doquet et Pilorgé, 2020). Ils constituent dans ce cadre-là une partie du geste professionnel de correction des copies : l'enseignant reçoit un devoir, le corrige et émaille ledit devoir de séries de remarques, sous des formes diverses, c'est-à-dire aussi bien verbales que graphiques (par exemple en soulignant certains mots, en reliant des passages à l'aide de flèches, etc).

Pour Halté (1984), ces annotations dans le cadre scolaire forment la matière d'un « dialogue pédagogique », dont il constate que, malheureusement, il n'est guère réalisé. Au niveau universitaire, Leboeuf (1999) déplore la rareté, quand ce n'est pas la totale absence, de commentaires réellement « explicatifs », c'est-à-dire qui donneraient à l'étudiant de réelles clés de compréhension de ses erreurs et défaillances.

Car, que ce soit dans le cadre scolaire ou dans le cadre universitaire, un des enjeux des commentaires inscrits par l'enseignant est d'offrir à l'étudiant un retour sur son travail ou un cadrage des attendus. Autrement dit, l'enseignant fournit, par ses commentaires, un moyen pour l'étudiant d'améliorer ses écrits, soit dans le cadre d'un aller-retour avant l'envoi du travail finalisé (par exemple pour un mémoire de master), soit dans la perspective d'un devoir ultérieur.

Dans la lignée de travaux déjà menés (Jacques et Charles, 2018 ; Jacques, 2022), nous examinerons ici la façon dont l'enseignant-e construit l'interaction avec l'étudiant-e, par fichier interposé. Nous analyserons la façon dont les enseignant-e-s se positionnent : peut-on déterminer un 'style' lié au correcteur ? Ou un 'style' lié au type de devoir ? Nous avons en effet un corpus qui comporte différents correcteurs-enseignants, et différents types de devoirs.

Corpus et méthodologie

Corpus

Nos analyses portent sur une partie du corpus *Littéracie avancée* (Jacques et Rinck, 2017), un corpus constitué de productions « écologiques » d'étudiants, c'est-à-dire de productions destinées aux évaluations et validations des cursus. Les textes ont été recueillis essentiellement dans des cursus de sciences du langage ou de didactique du français. Ce sont aussi bien des mémoires de master, des travaux d'études et de recherche, que des dossiers élaborés « à la maison » et autres types de devoirs. Les productions ont été recueillies directement au format numérique : un texte / un fichier. Cette particularité nous a donné accès aux annotations enseignantes entrées grâce à la fonction Commentaire des traitements de texte. Cette dernière

permet de recueillir simultanément le commentaire lui-même et la partie du texte sur laquelle il porte, comme le montre la figure 1.

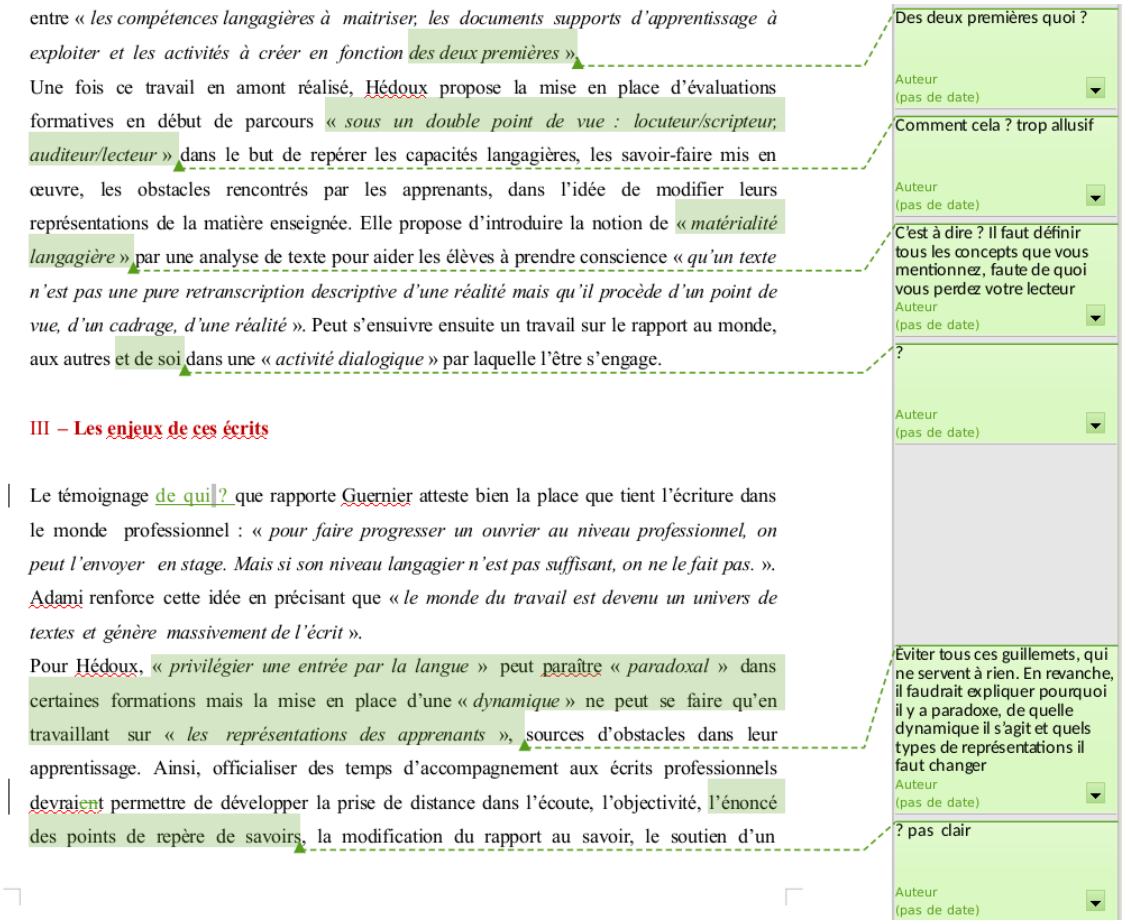


figure . 1 Aperçu de commentaires insérés

Seul un sous-ensemble de fichiers comporte de tels commentaires, environ 700 commentaires sont analysés.

Méthodologie

Nous nous intéressons ici à la façon dont les enseignants entrent en « discussion » avec l'étudiant auteur du devoir. Nous analysons donc les marques habituelles qui signalent les positions adoptées :

- d'éventuelles modalisations, telles que « peut-être » ou autre marque tendant à atténuer la position plus haute de l'enseignant,
- les temps verbaux, en particulier impératif / infinitif à valeur injonctive *vs* conditionnel,
- le lexique dans sa dimension axiologique.

L'analyse est semi-manuelle : certaines marques peuvent être repérées automatiquement, d'autres doivent faire l'objet d'une lecture individuelle. Ces marques seront mises en relation avec la nature du travail commenté et avec l'identité du correcteur afin de déterminer si l'une ou l'autre conditionne le positionnement du correcteur.

Nous nous intéressons aussi aux rares commentaires qui proviennent des étudiants eux-mêmes : ils insèrent à un point donné du mémoire une question pour laquelle ils attendent / espèrent un retour de l'enseignant.

Résultats

Nous ne pouvons ici avancer que des résultats préliminaires, les analyses n'étant pas toutes achevées.

Une première série d'investigations met en évidence une variété de positionnements. De la suggestion à la prescription, en passant par la proposition, les commentaires comportent plus ou moins d'atténuation, plus ou moins de marques de position haute.

Par exemple, en (1), l'utilisation de « un peu » tempère le qualificatif de « maladroit », même si ensuite c'est un impératif qui donne à l'étudiant un conseil de réécriture :

1. *Un peu maladroit; soyez plus sobre dans l'écriture*

De même, on peut constater un style globalement très injonctif, qui mêle impératifs et infinitifs (2), ou encore une modalité d'obligation (3) parfois atténuée par un conditionnel qui semble laisser le choix à l'auteur (4).

2. *Ne pas utiliser les en-têtes mais le corps du texte. Taper d'abord vos coordonnées, plus bas celles du destinataire. Utilisez ensuite les outils d'alignement: aligner à gauche vos coordonnées, à droite celles du destinataire.*

3. *Il faut choisir: soit vous indiquez les initiales, soit pas. Veillez à harmoniser votre mémoire*

4. *il faudrait peut-être donner une définition à cette expression*

De leur côté, les étudiants n'insèrent que très peu de commentaires – ce qui rend la généralisation hasardeuse – et le trait marquant de ces quelques adresses à leur enseignant est précisément qu'elles ne sont pas adressées. Là où les enseignants s'adressent directement à l'étudiant, à travers l'impératif notamment, l'étudiant s'adresse rarement à l'enseignant, l'exemple (5) est le seul qui interpelle directement l'enseignant.

5. *Ici Giasson est un auteur vulgarisateur, du coup pensez-vous que je peux la citer de cette façon ?*

Au contraire, la forme interrogative adoptée est relativement désincarnée :

6. Dois-je laisser ce passage ou bien est-ce que le précédent est suffisant ?

Ces interactions par fichier interposé exacerbent l'asymétrie de places enseignant / étudiant. La modalité injonctive parfois très forte positionne l'enseignant comme prescripteur et non comme conseiller.

Il nous reste à explorer la question de la relation entre ces caractéristiques et les devoirs annotés, et entre ces caractéristiques et les enseignants qui formulent les commentaires.

Références bibliographiques

Doquet, C., & Pilorgé, J.-L. (2020). La correction de copies au collège entre langue et discours : Une catégorisation syntactico-énonciative. *Repères*, *62*, 191-213. https://doi.org/10.4000/reperes.3274

Halté, J.-F. (1984). L'annotation des copies, variété ou base du dialogue pédagogique. *Pratiques*, *44*, 61-69.

Jacques, M.-P. (2022). Une typologie pour caractériser les commentaires des enseignants sur les textes des étudiants dans le supérieur. *Colloque Analyser de grands corpus scolaires et universitaires : des questions pour la recherche et pour la formation*, Bordeaux, juin 2022.

Jacques, M.-P., & Rinck, F. (2017). Un corpus de "littéracie avancée" : Résultat et point de départ. *Corpus*, *16*, 217-237. https://journals.openedition.org/corpus/2806

Jacques, M.-P., & Charles, É. (2018). Guider la réécriture dans l'enseignement supérieur : Analyse de pratiques enseignantes. *Le Français Aujourd'hui*, *203*, 135-146.

Leboeuf, P. (1999). « Très bien, Pascal, continue comme ça ! » ou la lecture de l'explication dans le commentaire. *Québec français*, *115*, 39-42. http://id.erudit.org/iderudit/56150ac

Réaliser un dictionnaire de prononciation à partir de corpus oraux

Gabriel Bergounioux [1] et Yvan Stroppa [1]
[1] Laboratoire LLL, Université d'ORLÉANS
gabriel.bergounioux@univ-orleans.fr, yvan.stroppa@univ-orleans.fr

Introduction

Comment réaliser en 2025 un dictionnaire de prononciation du français qui soit représentatif de la diversité des usages ? La disponibilité d'un nombre très important d'enregistrements en ligne résout la question de la disponibilité de données en masse, sans assurance de leur qualité ou de leur origine. Pour une exploitation contrôlée, doit-on privilégier l'extraction ou la génération, qu'elle soit produite à partir de textes ou dans une démarche participative comme les enregistrements du Wiktionnaire par exemple ? Quelles sont les opérations à prévoir dans la chaîne de traitement pour répondre aux attentes en matière d'analyse scientifique, d'applications didactiques ou industrielles et de diffusion auprès du public ? Comment construire une base de données évolutive qui intègre au fur et à mesure de nouvelles ressources ? Le programme « Passy » entend contribuer à l'ouverture d'une discussion à partir d'une maquette qui prélude à la réalisation d'un dictionnaire.

Corpus et méthodologie

Corpus

Le corpus utilisé est ESLO http://eslo.huma-num.fr/, plus particulièrement les entretiens (n = 83) d'ESLO2 (2007-2016), soit une centaine d'heures de dialogue entre un enseignant-chercheur en linguistique et un témoin sélectionné. La collecte ayant été réalisée à Orléans, la grande majorité des locuteurs sont issus de la région ou bien ils y résident depuis des décennies. Un échantillon de huit locuteurs a été retenu en contrastant les critères de genre, d'âge (+ ou – de 45 ans) et de catégorie sociale. Six locuteurs ayant un accent très perceptible ont été sélectionnés afin d'obtenir un contraste dans les réalisations : il s'agit de quatre « méridionaux » (Sud-Ouest et Nice), d'une personne originaire des Antilles et d'un Maghrébin vivant depuis longtemps en France.

Méthodologie

Une première liste de dix mots présentant des particularités phonologiques particulières (initiale en semi-voyelle (*huit*), hiatus (*dehors*)…) a été établie. Les occurrences ont été relevées dans les quatorze entretiens retenus et une transcription en API a été réalisée. L'étude ne porte pas sur la prononciation des mots, sur leur articulation, mais sur la forme telle qu'elle sera reconnue en contexte, c'est-à-dire sur l'extension des formes admises en audition par des locuteurs natifs.

Le sondage est destiné à définir les principes d'un dictionnaire en ligne où chaque entrée est associée :

- à la lemmatisation orthographique (base : le *TLFi*) – intégrant l'ensemble des formes fléchies ; - à la transcription phonétique « standard » et à l'identification de la catégorie grammaticale ;

- aux références normalisées dans des corpus libres de droit avec une indication chronométrique de la localisation dans l'enregistrement ;
- au lien vers la séquence sonore dans laquelle apparaît le mot ;
- aux métadonnées pour l'identification technique de l'enregistrement en EAD et pour décliner les propriétés sociales du locuteur.

En raison de leur présence massive dans le vocabulaire disponible, les noms propres, y compris les noms de marque, seront ajoutés autant que de besoin. en tenant compte de leur incidence sur l'acceptabilité phonotactique et la reconfiguration des correspondances graphème / phonème.

Résultats

Le programme est en phase de définition. L'objectif premier est d'ouvrir la discussion sur la façon de procéder pour la réalisation d'un site dont seront présentés l'affichage en lecture et la procédure d'enrichissement dans la perspective des sciences collaboratives.

Le projet s'articule à une reconstruction phonologique variationniste pour lequel PFC a rassemblé des données importantes. Il entend valoriser le travail sur corpus accompli par les chercheurs avec un contrôle des données et l'évaluation des résultats.

Références bibliographiques

Trésor de la langue française informatisé http://atilf.atilf.fr/

Wiktionnaire https://fr.wiktionary.org/wiki/Wiktionnaire

Bergounioux, G. (2016). Un essai de représentation de la forme sous-jacente. *Linguistique de corpus. Une étude de cas.* Paris, Champion, 73-100.

Côté, M.-H. (2019). Les prononciations du français : acquis et perspectives. *Cahiers de l'AIEF* 71, 13-36.

Durand, J., Laks, B., Lyche, C. (2005). Un corpus numérisé pour la phonologie du français (2005). *La linguistique de corpus*, Rennes, Presses Universitaires de Rennes, 205–217.

Martinet, A., Walter, H. (1973). *Dictionnaire de la prononciation française dans son usage réel*, Paris, France Expansion.

Retour vers le continu. Approche textométrique du liage dans des correspondances peu lettrées de la Première guerre mondiale

Agnès Steuckardt
[1]Praxiling, UMR 5267, Université de Montpellier Paul-Valéry
agnes.steuckardt@univ-montp3.fr

Introduction

L'émergence de la linguistique de corpus a vu la promotion d'une lecture discontinue du texte (De Angelis, 2017 : 90). Au début du 21e siècle, tout en assurant que son « propos n'est certes pas de « nier les propriétés séquentielles du texte », Jean-Marie Viprey revendiquait une approche outillée « non-séquentielle » et « tabulaire » des textes (2006 : 71) et Damon Mayaffre décrivait ainsi le changement d'approche induit par la linguistique de corpus : « les logiciels d'analyse de données textuelles commencent par faire exploser la linéarité du texte pour présenter leurs données en tableaux » et permettent une « lecture tabulaire et réticulaire », qui vient « en complément de la lecture linéaire usuelle » (2007 : 18-19). Viprey et Mayaffre marquaient alors une déférence polie à l'égard des travaux de linguistique textuelle, pour lesquels la nature linéaire du texte présente un caractère fondamental, puisque définitoire de leur objet.

Vingt ans plus tard, la linguistique de corpus poursuit son approche délinéarisée des textes. La taille des corpus a considérablement augmenté (Vanni, Mittmann, 2016 : 8), décourageant peut-être le retour aux textes qui les constituent et à leur lecture linéaire. Une place notable est faite bien sûr à l'analyse des séquences discursives récurrentes, qu'elle aborde par les notions de phraséologismes, motifs, routines ou encore formules (Steuckardt *et alii*, 2022 : 1-2) ; il s'agit cependant de séquences relativement brèves, qui segmentent le discours en une série de fragments : leur structuration globale à l'échelle d'ensembles discursifs étendus ou de textes semble demeurer quelque peu hors du champ d'intérêt de la linguistique de corpus. L'approche textométrique, dont l'entité de base est la chaîne de caractères entre deux blancs, donc discontinue, est-elle par nature réfractaire à l'appréhension de la continuité textuelle ?

Les procédures de la continuité textuelle ont fait l'objet de théorisation dans le champ de la grammaire et de la linguistique textuelles. Jean-Michel Adam a notamment cherché à décrire « la fabrique du continu » du mot à la proposition, de la proposition à la phrase, de la phrase au texte, en intégrant les niveaux sémantiques, énonciatifs, pragmatiques (Adam, 2020 : 56).

On tentera ici une exploration du liage assistée par TXM, en s'appuyant sur un corpus qui nous semble le questionner de façon récurrente et radicale. La question du liage se pose en effet en deçà même du niveau lexical : lier les lettres est un problème pour les scripteurs peu lettrés de Corpus 14, un corpus de 2171 lettres de familles peu lettrées, écrites pendant la Première Guerre

mondiale (Steuckardt, Luxardo, 2024, version 3)[1]. On propose d'utiliser l'outil textométrique pour parcourir dans une première partie les niveaux du liage, de la lettre à la séquence lexicale, et dans un second temps de passer au liage textuel, en développant une étude de cas avec des ligateurs énonciatifs caractéristiques de ce corpus, *tu me dis* et *je te dirai*.

Corpus 14, un révélateur du liage textuel

Constitué de correspondances familiales de la Première Guerre mondiale, Corpus 14 met en évidence le problème qui se pose à nous quand il faut assembler entre eux des éléments discursifs : pour les scripteurs et scriptrices peu lettrés de Corpus 14, peu habitués à l'écriture mais contraints par la guerre à la pratiquer, la question du liage se pose à tous les niveaux de la production du discours écrit, comme on peut le percevoir par exemple dans ce début de lettre :

> Je fais reponse a vaux aimable lettre que j'ai reçue avec plaisir sur tout en na prenand que vous éte tous en bonne santée car il la n'est de méme pour moi : Tu me di que tu ne peu pas me donner des nouvelles de nainé inci que d'Auguste ébien ja n'est eu par et l'Eonore (Laurent, 16/02/1916)

On prendra cet extrait comme point d'appui pour poser la question du liage à différents niveaux linguistiques.

Le liage graphique

Au niveau graphique, comment lier les lettres pour former les mots ? Ce qui pour les lettrés est un automatisme est, pour ces scripteurs peu expérimentés, un chemin à imaginer. Les frontières séparant les entités lexicales ne s'imposent pas toujours à eux : du point de vue du standard, leurs graphies sont analysées comme des sur-segmentations, comme les séquences *sur tout* et *et l'Eonore* de l'échantillon cité (Dal Bo, 2024), ou des sous-segmentations comme *ébien* (Surcouf, 2024 : 299). Arrêtons-nous dans cette étude consacrée au liage, sur un exemple de ces sous-segmentations, qui d'après l'étude exploratoire de Beatrice Dal Bo représentent 82% des cas de découpage non standard.

Les liaisons de l'oral viennent interférer avec la perception des frontières lexicales, comme le montre la séquence *en na prenand :* la liaison entre le marqueur *en* et le gérondif *apprenant*, produisant une séquence phonétique [ɑ̃napRənɑ̃], déclenche un liage graphique [na]. La segmentation se fait alors en fonction des mots graphiques connus du scripteur : *en* et *prenant*, participe de *prendre* : il lui reste ce *na*, dont il fait l'hypothèse qu'il s'agit d'un marqueur autonome, hypothèse lexicale attestée 526 fois dans Corpus 14, et à laquelle se rangent 16 scripteurs différents. Le tri des concordances permet de mettre en évidence les différents usages de ce *na*. Il est utilisé d'une part pour traiter la liaison *en* (160 occurrences), *a* [pour *en*] (17) ou *on* (94) suivi un mot à initiale vocalique, soit 271 occurrences :

> tout le monde a na par dessur les oreilles (Laurent, 12/12/1915)
>
> on na encore 2 morceaux de lart (Anne-Marie, 29/04/1916)
>
> ont na roullé et hersé les grains tout ces jours ci (Anne-Marie, 25/04/1917)

Il sert d'autre part à traiter les séquences comportant le discordanciel *ne* élidé (255) :

> le capitaine na pas voulut (Alfred, 17/11/1914)

---

1 La version 3 permet d'interroger le corpus dans la transcription fidèle, qui respecte les graphies originales des lettres, ou dans la transcription orthographiée, plus adaptée à l'exploration textométrique.

tu na peut être pas tant d'amitiée pour moi (Victoria, 13/01/1916)

La sous-segmentation *na* suggère d'étendre les requêtes sur les liaisons : en poursuivant sur la liaison [n] suivi d'un mot à initiale « a », on relève ainsi en trois lettres nai/nan/nas (171 occurrences), en 4 lettres nais/nait/navu (26), et ainsi de suite. Une étude systématique de la sous-segmentation, prenant appui sur les études exploratoires de Dal Bo et Surcouf et sur les fonctionnalités de TXM permettrait ainsi de quantifier et donc de mieux décrire les faits de liage graphique.

Le liage lexical et phraséologique

Dans les lettres des scripteurs, comme plus généralement dans le discours ordinaire, les mots des scripteurs s'enchaînent selon des séquences relativement prévisibles. Les ouvertures et clôtures de lettres en particulier présentent un caractère formulaire qui a été souvent remarqué. Par exemple, le début de la lettre de Laurent constitue un motif à 13 mots. La séquence est un départ de lettre partagé par des scripteurs d'autant plus nombreux qu'on réduit la taille du segment :

Séquence (transcription orthographique)	Nombre de scripteurs	Nombre d'occurrences
Je fais réponse à votre/vos aimable(s) lettre(s) que j'ai reçue(s) avec plaisir	1	26
Je fais réponse à votre aimable lettre	2	53
Je fais réponse à votre	3	81
Je fais réponse à	6	238

table 1. : ***Je fais réponse à* dans Corpus 14**

Au niveau phraséologique, l'analyse des cooccurrences proches (avec une distance contextuelle inférieure à 1 mot) montre que la locution verbale *faire réponse* est la deuxième en fréquence des locutions en *faire*, la première étant *faire plaisir*. L'analyse des cooccurrences plus lointaines (avec une distance contextuelle inférieure à 9 mots) met en évidence une autre locution verbale remarquable en *faire*, la locution *se faire du mauvais sang* :

Occ	Fréq	CoFréq	Indice	DistMoy
sang	488	372	167,502	2,9435484
mauvais	692	458	165,0035	2,1419213
réponse	495	360	150,693	0,25277779
plaisir	1434	687	139,584	0,93449783

table 2. : Cooccurrences de *faire* dans Corpus 14

La locution est très majoritairement employée à la forme pronominale négative. La requête

[frlemma="faire"][word=".*"][word=".*"][word="mauvais"][word="sang"]

ramène 274 occurrences, l'énoncé prototypique étant *ne te fais pas du mauvais sang* (190 occurrences)22. L'outil textométrique est particulièrement adapté au repérage de ces liages lexicaux, qui sont le départ de blocs discursifs généralement désignés sous l'appellation de *motifs*.

2 La forme plus soutenue, qui veut que le partitif se réduise à de dans les énoncés négatifs (Grévisse, 1988 : §569), ne te fais pas de mauvais sang ne présente que 11 occurrences.

Le liage syntaxique

Mais comment s'enchaînent entre eux ces motifs ? Au niveau syntaxique, comment les scripteurs passent-ils d'une séquence à une autre ? Il n'y a pas, comme le soulignent naguère l'approche du lexique-grammaire défendue par Gaston Gross ou celle de la grammaire de construction, de véritable solution de continuité entre le niveau lexical et le niveau syntaxique : les séquences lexicales repérées dans la sous-partie précédente se poursuivent au-delà des limites grammaticales de la proposition, le pronom relatif *que* liant la proposition inaugurale à la suivante, qu'elle lui subordonne. Une approche plus spécifiquement syntaxique du liage peut cependant se centrer sur les items identifiés comme des connecteurs. Ainsi notre extrait atteste l'usage de car pour enchaîner les propositions et passer des nouvelles reçues (« vaux aimable lettre que j'ai reçue avec plaisir sur tout en na prenand que vous éte tous en bonne santée ») aux nouvelles données (« il la n'est de méme pour moi ») ?

Avec 4400 occurrences dans Corpus 14, la conjonction *car* occupe le deuxième rang des conjonctions de coordination, après et (12533 occurrences). Devançant les conjonctions *mais* (3819) et *ou* (724), elle apparaît une ressource appréciée par les peu lettrés, même si le lien causal n'est pas toujours une évidence : en quoi ici le fait d'être en bonne santé est la cause de la joie éprouvée à réception des bonnes nouvelles des siens ? Dans ces lettres où la ponctuation est rare, le connecteur *car*, bref et graphiquement commode, vient marquer les enchaînements de la pensée, préférentiellement à *parce que* (52 occurrences). Cette utilisation assez spécifique des connecteurs dans l'écrit peu lettré pourrait elle aussi donner lieu à une exploration systématique.

On a ici proposé d'appréhender la notion de « continu » plus largement que ne le fait la linguistique textuelle : en partant de la lettre et le phonème, pour aller au morphème, au lexème et à ce qu'il régit dans la proposition, et à l'enchaînement syntaxique. L'outil textométrique, par ses fonctionnalités de repérage des concordances et cooccurrences, donne accès à un repérage du continu à ces différents niveaux. Mais peut-il aussi soutenir l'appréhension du continu au niveau textuel ?

TXM : un outil pour l'exploration des ligateurs textuels

La linguistique textuelle envisage des types de liage se situant sur différents plans, principalement sémantique (anaphores, isotopies), macro-syntaxique et énonciatif (connecteurs), pragmatique (implicite, chaînes d'actes de discours) (Adam, 2020 : 123-89 ; Caillat, Verine (dir.), 2024 : 39-41). Si le plan pragmatique paraît peu accessible à l'exploration textométrique, l'inscription des deux autres plans dans le matériau linguistique donne des prises pour appréhender le liage textuel. Ainsi, les travaux de Traitement Automatique des Langues, en cherchant à repérer les chaînes de référence, à travers l'annotation des expressions référentielles et des anaphores (Landragin, 2020), en repèrent une des procédures, même si leur objet de recherche est l'entité nommée et les différentes manières dont elle l'est, plutôt que le liage lui-même. Comment procèdent les scripteurs peu lettrés pour réaliser le liage textuel de leur lettre ?

Le liage textuel dans les lettres substandard

Comme dans d'autres genres, le liage textuel peut s'opérer dans le plan sémantique. Notre échantillon permettra d'illustrer le tissage opéré par les anaphores (en gras) et l'isotopie de la lettre (souligné) :

> Je fais réponse à vos aimables lettres **que** j'ai reçues avec plaisir surtout en apprenant que vous êtes tous en bonne santé car il **en** est de même pour moi : Tu me dis que tu ne peux pas me donner des nouvelles de Néné ainsi que d'Auguste, eh bien j'**en** ai eu par Éléonore

Ces procédures contribuent à assurer la cohésion textuelle, sans être spécifiques au genre épistolaire.

Sur le plan macro-syntaxique et énonciatif, le genre épistolaire présente davantage de spécificité. Les descriptions des correspondances « substandard » soulignent en effet la forte armature textuelle qui cadre la lettre (Bruneton-Governatori, Moreux, 1997 ; Rutten, Van der Wal, 2014) ; elle s'inscrit dans une structure prévoyant trois parties : la formule d'ouverture, le corps de la lettre, la formule de clôture. Les formules d'ouverture et de clôture sont elles-mêmes subdivisées respectivement en lieu, date, salutation, formule intersubjective d'une part et salutation, signature d'autre part. Cette architecture préconstruite autorise le passage d'une partie à l'autre sans transition : on a pu observer par exemple que, dans 65% des cas, il y a absence de transition entre la formule d'ouverture et le corps de la lettre (Grosse *et alii*, 2016).

Il n'en reste pas moins que, dans plus d'un tiers des cas, les scripteurs ont eu recours à un « liage textuel » entre la formule d'ouverture et le corps de la lettre. Il peut s'agir d'un connecteur tel que *et*, *car*, ou *mais*, comme par exemple dans :

> je suis toujour en bonne santée mais nous avons quittér Maixe hier a Midi (Alfred, 22/10/1914)

Mais plus généralement le scripteur passe par un liage à la fois syntaxique et énonciatif, enchaînant à la formule intersubjective un verbe déclaratif en position de circonstant, comme dans la lettre d'André, ou d'épithète, comme dans celle de Pierre :

> Chers parents et cher Louise je vous envoie ces quelques lignes pour dire que je suis toujours en bonne santé et je crois que vous etes tous de même (André, 07/08/14),
>
> Chère Epouse J'ai reçue ta lettre ce matin me disant que mon cousin de campdeltour a ecrit (Pierre, 05/11/1914).

La lettre échantillon de Laurent n'utilise ni connecteur standard, ni rection syntaxique pour passer de la formule d'ouverture au corps de la lettre : on pourrait catégoriser ce cas comme une absence de transition. Pourtant, l'usage de *Tu me dis* manifeste un passage au niveau énonciatif analogue à ce que l'on a observé dans les deux derniers exemples. Ne peut-il pas fonctionner comme un ligateur textuel ?

Des ligateurs énonciatifs : *tu me dis*, *tu me diras*, *je te dirai*

Si l'on envisage la continuité syntaxique, *tu me dis* se trouve en parataxe par rapport à la séquence précédente : à la différence du complément prépositionnel à l'infinitif ou du participe épithète, il n'existe pas de rapport de dépendance avec ce qui précède. Toutefois, ce *tu me dis* introduit une continuité analogue à celle de l'épithète *me disant* de Pierre ; Laurent aurait pu écrire :

> Je fais réponse à vos aimables lettres, que j'ai reçues avec plaisir surtout en apprenant que vous êtes tous en bonne santé car il en est de même pour moi, me disant que tu ne peux pas me donner des nouvelles de Néné

La série enchâssée qui commence à *que j'ai reçues* et va jusqu'à *pour moi* engage le scripteur à préférer une parataxe, mais le lien demeure entre le mot *lettres* et *Tu me dis*, c'est-à-dire entre une des lettres et son contenu. Le ligateur opère le passage entre le mot *lettre*, qui porte un sème

énonciatif, et l'énoncé rapporté *que tu ne peux pas me donner des nouvelles de Néné*. Sur le plan syntaxique, s'il est en rupture avec le contexte avant, il fonctionne comme un introducteur du contexte après. Employé par 15 scripteurs différents, il présente 833 occurrences dans le corpus. Son statut de ligateur serait à évaluer plus précisément en fonction de sa situation dans la structure de la lettre : on se heurte là à une limite de l'exploration textométrique ; cependant, les sondages effectués mettent en évidence une situation de pivot. *Tu me dis* est un moyen habituel de passer de la formule d'ouverture au corps de la lettre :

> Aujourd'hui j'ai reçu deux lettres l'une du 22 et l'autre du 23, cela me contente beaucoup quand je reçois tes chères nouvelles. Pour le moment je suis en très bonne santé et souhaite que tu en sois de même. Tu me dis que mon frère Hippolyte est pas venu vous voir (Félicien, 26/08/1915)

> Je viens de recevoir ta carte avec beaucoup de plaisir, surtout de te savoir en bonne santé, pour moi il en est de même. Tu me dis qu'encore tu ne sais pas le jour que tu viendras, mon Dieu que ce jour se fait désirer. (Victoria, 22/12/1916)

Les scripteurs y ont également recours dans le corps même de la lettre, pour passer d'une thématique à une autre :

> Tu peux croire qu'en ce moment-ci, il ne fait pas chaud, jamais je n'avais vu un froid pareil. Tu me dis que Charles est arrivé en permission, je l'ai manqué à Perrache (Félicien, 24/01/1917)

> Il a passé quelques jours qu'il a fait un froid terrible, nous avons un peu de la neige, maintenant il fait un peu meilleur, mais il gèle toujours, les nuits sont bien froides et sûrement qu'au pays il en doit être la même chose, surtout avec la neige que vous avez. Tu me dis sur ta lettre que tu as pas beaucoup du bois, j'écrirai à mon frère qu'il ira tout tomber. (Jules, 07/01/1918)

À l'aide des expressions régulières, on peut en repérer les variantes de ce *tu me dis*. La requête

> [frpos="PRO:PER"][frpos="PRO:PER"][frlemma="dire"]

ramène 2708 occurrences, dont une partie ne relève pas directement du liage. Ainsi, dans les 212 formes de 3^{e} personne, le marquage de continuité s'opère par le pronom anaphorique plutôt que par l'introducteur déclaratif et ne se situent pas, d'après les sondages effectués, dans des situations charnières. De même, les formes de passé (143 au passé composé, 93 à l'imparfait) sont généralement intégrées dans des subordonnées, notamment en *comme*, ainsi dans :

> Vivons toujours dans l'espoir, comme je te disais dans ma dernière lettre (Félicien, 20/04/1915)

Prise dans la subordonnée *comme je te disais*, la séquence en *dire* ne relève pas ici du liage mais de la modalisation autonymique.

Les formes de futur sont plus abondamment représentées, avec 348 occurrences de *tu me diras*, et surtout 512 de *je te dirai*. Fonctionnent-ils comme ligateurs ? La séquence *tu me diras* se distingue de *tu me dis* en ce qu'elle n'est pas directement ancrée dans le contexte avant : contrairement à *tu me dis*, elle ne se réfère pas explicitement à un mot du type *lettre*, ni implicitement à la lettre reçue. Dans 221 occurrences, soit 63,5% des cas, elle sert à introduire une interrogative indirecte, comme dans :

> Il y a eu aussi deux porte-plumes très jolis et trois autres faits avec des cartouches boches. Tu me diras, cher époux, si pour venir à toi, les lettres y mettent tant de temps comme pour nous ici (Marie, 13/05/1915)

Tu me diras introduit ici un nouveau thème, donc marque une rupture avec ce qui précède, mais il le fait en explicitant le fil énonciatif de la lettre : cette explicitation introduit une forme de continuité textuelle, tenant au rappel de la trame énonciative qui sous-tend le texte épistolaire. Alors que *tu me dis* assure une continuité textuelle à la fois sur le plan sémantique et sur le plan énonciatif, *tu me diras* ne fonctionne comme ligateur que sur le plan énonciatif : on peut le considérer comme un ligateur textuel plus faible que *tu me dis*.

L'emploi de *je te dirai* peut surprendre au premier abord. À la différence de *tu me diras*, la séquence *je te dirai* n'est pas là pour amener une question : elle n'est jamais suivie par un *si ;* dans 412 occurrences sur 501, soit 81,6%, elle introduit un *que*, comme dans :

> Je fais réponse à ton aimable lettre datée du 25 laquelle m'a fait plaisir surtout en apprenant que vous êtes tous en bonne santé car il en est de même pour moi. Je te dirai que j'ai reçu l'imperméable hier 29, il est arrivé en bon état. (Laurent, 30/12/1915)
>
> Je t'écris ces 2 mots à la hâte pendant les 5 minutes qu'on a de repos ; je te dirai que j'ai changé de bataillon comme je t'ai déjà marqué sur l'autre lettre. (Henri, 28/11/1914)
>
> La neige est tombée, aujourd'hui lundi il a en a plus. Il gêle. Je te dirai que j'ai retrouvé ma ceinture. (Joseph, 21/02/1916)

À quoi sert d'introduire ce que l'on va dire en disant qu'on va le dire ? Alors que *tu me diras* est utile pour déplacer la scène énonciative dans le futur éloigné de la réponse, encore inconnue, le déplacement qu'accomplit *je te dirai* ne semble pas nécessaire : pourquoi déplacer lourdement la scène énonciative vers le futur alors que le propos sera prononcé et connu à l'instant immédiatement consécutif à l'énonciation de *je te dirai* ? On peut penser que ce *je te dirai* est formé par analogie avec *tu me diras* : s'il n'a pas de réelle utilité dans la représentation temporelle, il explicite, comme lui, la trame énonciative qui sous-tend la lettre. Ainsi, les séquences *tu me dis*, *tu me diras*, *je te dirai*, en rappelant le contexte énonciatif et sa continuité permettent d'atténuer la rupture thématique, et font office de ligateur textuel.

Conclusion

Enchaîner lettre après lettre, mot après mot, séquence après séquence n'a rien d'une évidence pour les scripteurs peu lettrés, c'est pourquoi Corpus 14 permet, sans doute mieux que d'autres corpus à l'écriture plus fluide, de saisir les lieux où achoppe la continuité textuelle. Il invite à la poser dès le niveau graphique, pour arriver jusqu'au niveau du texte, sur lequel se focalise traditionnellement la linguistique textuelle. On a envisagé ici la notion de continuité textuelle de façon plus large que ne le fait habituellement cette discipline, en partant des opérations de liage réalisées en amont.

Au niveau graphique, l'approche textométrique fait apparaître les continuités phonétiques qui, notamment par le biais de formes telles que *na*, *ya* ou *quil*, se glissent dans l'écrit peu lettré. Au niveau lexical, elle permet, notamment par les requêtes à expressions régulières, de repérer les motifs, qui amorcent l'enchaînement menant du mot vers la séquence discursive. Au niveau syntaxique, elle peut assister le repérage des connecteurs qui lient entre elles ces séquences. Les aspects sémantique, énonciatif et pragmatique de la continuité textuelle offrent moins de prise à l'exploration textométrique. Les recherches sur les anaphores ou sur les isotopies (Valette, 2010) ont ouvert des pistes dans cette perspective ; le cas des ligateurs *tu me dis*, *tu me diras*, *je te dirai*, repérés dans Corpus 14, invite à envisager le fil énonciatif en tant que vecteur de continuité.

Un élargissement de l'enquête pourra porter sur la nature du corpus interrogé : un corpus tel que LesVocaux (Glikman *et alii*, 2025) présente des convergences de caractéristiques avec Corpus 14, et semblerait aussi une ressource intéressante pour l'exploration du liage textuel. On y relève notamment, comme dans Corpus 14, un usage *tu* + *dire* (39 occurrences) et *je* + *dire* (66 occurrences) en tant que ligateurs, mais avec des formes, proportions et nuances différentes, sur lesquelles une approche contrastive pourrait être développée.

Références bibliographiques

Adam, J.-M. ([2005] 2020). La linguistique textuelle : Introduction à l'analyse textuelle des discours, Paris, Armand Colin.

Bruneton-Governatori, A., Moreux, B. (1997). Un modèle épistolaire populaire. D. Fabre (dir.), *Par écrit. Ethnologie des pratiques d'écriture quotidiennes*, Paris, Éditions de la Maison des Sciences de l'Homme, 79-103.

Caillat, D., Verine, B. (dir.) (2024). *Manuel d'analyse du discours*, Louvain-la-Neuve, De Boeck.

Dal Bo, B. (2024) ; Une étude de la sur-segmentation des mots graphiques dans des écrits de scripteurs peu lettrés. *Le Français moderne*, 92, en ligne.

De Angelis, Rossana (2017). La linguistique de corpus à l'épreuve du numérique : textes, textures, documents. *Dossiers d'HEL*, *Analyse et exploitation des données de corpus linguistiques*, 11, 81-95.

Glikman, J., Mazziotta, N., Benzitoun, C., et al. (2025). *LesVocaux* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v0.0.2, https://hdl.handle.net/11403/lesvocaux/v0.0.2.

Grévisse, M., Goose, A. (1988). *Le bon usage*, Paris et Louvain-La-Neuve, Duculot.

Grosse, S., Steuckardt, A., Sowada, L., Dal Bo B. (2016). Du rituel à l'individuel dans les correspondances peu lettrées de la Grande Guerre, F. Neveu *et alii* (éds), *Actes du 4e Congrès mondial de linguistique française*, EPD Sciences, 1-15. DOI 10.1051/shsconf/20162706008

Landragin, F., *Democrat : description et modélisation des chaînes de référence, outils pour l'annotation de corpus (en diachronie et en langues comparées) et le traitement automatique.* Rapport final du projet validé par l'ANR, 2020.

Mayaffre, D. (2007). Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques. F. Rastier, M. Ballabriga, (éd.), *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation*, Toulouse, PUT, 15-26.

Rutten, G., van der Wal, M. J. (2014). *Letters as Loot. A sociolinguistic approach to seventeenth- and eighteenth-century Dutch*. Amsterdam / Philadelphia : Benjamins.

Steuckardt, A., Grosse, S., Dal Bo, B., Sowada, L. (2022). La routine et le style. Exploration outillée des formules d'ouverture et de clôture dans des correspondances peu lettrées de la Première Guerre mondiale. Idmhand F., Marasescu-Galleron I. (dir.), *Dix ans de corpus d'auteurs*, France, Éditions des archives contemporaines, 203-220.

https://eac.ac/publications/9782813004352

Steuckardt, A., Luxardo, G. (2024). Corpus 14, version 3, https://textometrie.univ-montp3.fr/txm/ (version 2 Praxiling - UMR 5267 (2019). *Corpus 14* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v2, https://hdl.handle.net/11403/corpus14/v2).

Surcouf, C. (2024). L'efficacité orthographique des peu-lettrés » : une analyse des graphies des Poilus du Corpus 14. Steuckardt A., Gomila C., Wionet C. (dir.), *Gens ordinaires dans la Grande Guerre*, Paris, Éditions de la Maison des Sciences de l'Homme, 285-304.

Vanni, L., Mittmann, A. (2016). Cooccurrences spécifiques et représentations graphiques, le nouveau « Thème » d'Hyperbase. *JADT 2016 - Statistical Analysis of Textual Data*, Jun Nice, France, 295-305.

Valette, M. (2010). Propositions pour une lexicologie textuelle. P. Blumenthal, S. Mejri (dir.), *Les configurations du sens*, *Zeitschrift für Französische Sprache und Literatur*, 37, 171-188.

Viprey, J.-M. (2006). Structure non-séquentielle des textes, *Langages*, 163, 71-85.

Réunion de travail : comment un participant allophone peut montrer qu'il n'est pas certain d'avoir bien compris ?

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

Au cours de leur stage, dans leurs activités sportives ou culturelles ou un peu plus tard dans leur entrée dans la vie professionnelle, les allophones sont assez rapidement conviés à des réunions, quelles que soient leur profession ou leurs activités (Borzeix *et al.*, 2001). Suivant s'il s'agit de réunions régulières d'avancement, de réunions exceptionnelles pour discuter d'un projet particulier ou bien de l'assemblée générale d'une association, leur format et leur organisation peuvent changer mais certaines étapes-clés se retrouvent dans leur déroulement comme l'ouverture et la clôture de la réunion, les changements de sujet, la prise de parole, les demandes d'informations ou de précisions, les évaluations, les marques d'accord et de désaccord, les négociations ou encore la prise de décision. Au cours de ces réunions, des informations importantes seront échangées dont les participants allophones doivent avoir connaissance ou bien des décisions seront prises concernant les tâches qu'ils doivent effectuer, par conséquent il est important qu'ils comprennent les échanges et leur implication dans l'organisation de leur travail (Lacoste 2001). Pour y parvenir, nous nous proposons de montrer comment les natifs eux-mêmes vérifient leur compréhension ou demandent des précisions *in situ* dans des réunions professionnelles écologiques recueillies dans différents contextes.

Ressources et méthodologie

Cette communication découle du projet Interfare, Interagir plus facilement en réunion de travail, où les interactionnistes des laboratoires ICAR et ATILF ont transposé les résultats des analyses interactionnelles des réunions professionnelles en ressources pour des publics allophones, mais également des publics débutant dans la vie active ou en reconversion professionnelle (https://icar.cnrs.fr/interfare/). Le site s'appuie sur un découpage des réunions en une quinzaine d'étapes-clés (Ravazzolo *et al.*, 2023) faisant l'objet d'une description détaillée qui met en perspective les principales fonctions langagières avec la variété des procédés, verbaux et multimodaux, permettant de les réaliser. Par exemple, le rôle des regards de l'animateur adressés à l'ensemble des participants ou des pauses avant de changer de sujet.

Notre démarche s'inscrit dans une perspective interactionnelle où les contributions des différents locuteurs sont prises en compte afin de mener à bien leur échange, en étudiant la manière dont ils se coordonnent et ajustent continuellement leur propos aux réactions de leurs interlocuteurs « le locuteur adapte son tour de parole à son ou ses interlocuteurs au fur et à mesure de sa production » (Traverso, 2016 : 25). Dans le cadre d'une réunion de travail, la présentation d'une solution exposée par un collègue projette par exemple une évaluation de celle-ci par les autres participants qui conduira à son acceptation, à un rejet ou bien à une

alternative suivant les retours qu'elle suscite (Bonu 2001). Parmi les corpus de réunion, nous avons donc sélectionné différents extraits qui représentent les procédés les plus fréquemment observés *in situ* pour réaliser une activité donnée, en présentant plusieurs variantes possibles que nous avons explicitées (Traverso 2011). En effet, cette variation permet de s'adapter au contexte et caractérise une interaction il s'avère donc pertinent de la présenter à un participant allophone afin de lui donner les moyens de la comprendre et de la reconnaitre quand il y sera confronté (Alberdi & Etienne, 2023).

À ces 200 extraits contextualisés, transcrits et décrits étape par étape, viennent s'ajouter une centaine d'activités pédagogiques multimédia qui permettent de travailler les spécificités multimodales en situation de réunion (Mondada 2008), qu'elles concernent la prosodie (intonations, saillances…), les gestes (pointage, postures, hochements de tête ou d'épaules, manipulation d'artefacts…), les regards, vers un locuteur donné pour le désigner ou vers l'ensemble des participants, ou encore l'organisation de la parole (réponses chorales, aménagement de pauses, chevauchements de parole…). D'autres exercices portent sur des expressions figées fréquentes en réunion ou sur certaines caractéristiques de l'oral comme les répétitions, les demandes de confirmation ou certains formats de questions afin d'en étudier les finalités.

En parallèle de ces activités de compréhension, des scénarios sont proposés pour encourager la production d'après ce qui a été étudié, en donnant des consignes précises et en allouant des rôles variés aux locuteurs dont le nombre varie en fonction du contexte.

Orientée vers les enseignants de formation de langue ou de formation professionnelle, cette plateforme Interfare permet également de cibler un sous-ensemble de ressources avec l'onglet « Pour commencer » qui restreint le nombre d'extraits et d'exercices par rubrique afin de s'adapter à des modules d'enseignement plus court, voire de sensibilisation au contexte particulier de la réunion, sans pour autant sacrifier la variation et préconiser une seule manière de procéder.

Méthodologie

Dans cette communication, nous nous proposons d'exposer l'apprenant à différents mécanismes qui lui permettront de vérifier sa compréhension, de demander davantage d'informations ou encore de faire préciser le sens d'un terme qu'il ne parvient pas à comprendre en contexte. Un allophone est toujours réticent à montrer son incompréhension, et suppose qu'il pourra récupérer le fil de la discussion un peu plus tard alors qu'il court au contraire le risque de laisser échapper une information importante, de perdre le sujet en cours de discussion ou la possibilité de se positionner sur une solution qui le concerne, simplement parce qu'il n'a pas compris une notion qui aurait pu être rapidement éclaircie. Dans cette perspective, il nous a semblé intéressant de montrer comment les participants natifs eux-mêmes interviennent pour demander des explications ou vérifier leur compréhension, alors qu'ils ne se rencontrent a priori pas de problèmes de langue, afin de montrer à quel moment intervenir et de quelle manière. Cette exposition à des difficultés authentiques dans plusieurs contextes de réunion, avec des participants de niveau hiérarchique différent, permet de démystifier la situation de réunion qui parait souvent formelle et figée aux allophones, les rendant réticents à en perturber le déroulement, alors qu'elle présente un caractère adaptatif comme dans toute autre interaction « les pratiques sociales et langagières mobiliseraient des manières de faire et d'agir en partie stabilisées, qui ont toujours un côté contingent, adapté, flexible. » (Pekarek-Doehler 2006 : 32).

Résultats

Nous nous appuierons donc sur plusieurs rubriques d'Interfare dont l'étape-clé « Mieux s'exprimer » et les fonctions « Compléter » et « Questionner » de la rubrique « Prendre la Parole » afin de construire une séquence pédagogique complète sur les différentes possibilités d'intervention d'un participant allophone rencontrant des difficultés de compréhension. Cette séquence sélectionnera et articulera plusieurs activités pédagogiques multimédia parmi celles disponibles afin de montrer comment travailler certains procédés verbaux (répétitions, reformulations, marqueurs…) et multimodaux (gestes, regards, changements de posrture…) en contexte.

Références bibliographiques

Alberdi, C., Etienne, C. 2023. Aider l'apprenant à parler en interaction, du coup comment plaisanter ou refuser ? Repères Dorif 28. Entre le théorique et l'expérentiel : l'oral en didactique du FLE. Questionnements et perspectives.

Bonu, B. (2001). Les évaluations conversationnelles dans la narration. Revue québécoise de linguistique, 29(1), 51–69. https://doi.org/10.7202/039429ar

Borzeix, A., Fraenkel, B., & Boutet, J. (2001). Langage et travail : Communication, cognition, action. CNRS éditions.

Lacoste, M. (2001). Peut-on travailler sans communiquer ? Dans Borzeix, A. et Fraenkel,

Mondada, L. (2008). Production du savoir et interactions multimodales : Une étude de la modélisation spatiale comme activité pratique située et incarnée.

Pekarek-Doehler, S. (2006). Compétence et langage en action ». Bulletin Suisse de Linguistique Appliquée, 84, 9-45.

Ravazzolo, E., Etienne, C., André, V. (2023). Comment enseigner l'oral en classe ? L'exemple d'un dispositif pour apprendre à interagir plus facilement en réunion de travail. Repères Dorif 28. Entre le théorique et l'expérentiel : l'oral en didactique du FLE. Questionnements et perspectives.

Traverso, V. (2011). Analyser un corpus de langue parlée en interaction : Questions méthodologiques. Verbum : Analecta Neolatina, 2011, 4, 313.

Traverso, V. (2016). Décrire le français parlé en interaction. Éditions Ophrys.

Terrain, corpus, typologie : de la collecte à la réutilisation des données

Léa MOUTON[1], Karl SEIFEN[2] et Alice VITTRANT[3]
[1]CNRS, Laboratoire Interactions, Corpus, Apprentissages, Représentations (ICAR)
[2]Aix-Marseille Université, Laboratoire Parole et Langage (LPL)
[3]Université Lumière Lyon2, Laboratoire Dynamique Du Langage (DDL)
lea.mouton@cnrs.fr, karl.seifen@univ-amu.fr, alice.vittrant@univ-lyon2.fr

Introduction

Notre présentation sera consacrée aux données de terrain, à leur constitution en corpus annotés, à leur mise à disposition pour les communautés de spécialistes, pour permettre leur réutilisation. En d'autres termes, nous suivront le cycle de vie des données langagières en linguistique descriptive et typologique.

Les données collectées par les auteurs sont de différents types : données de langue écrite sans standardisation, données de langue orale non-décrite, données dialectales. Ces types de données sont parfois accessibles sur des sites dédiés aux langues rares (Pangloss), aux langues orales (CoCoON) (Jacobson *et al* 2015) ou liés à des aires géographiques particulières (AILLA). Cet effort vers la science ouverte, parfois difficile avec les données de terrain (Michaud & Noûs 2024), peut permettre la réutilisation de corpus à des fins de recherche typologique, comme illustré par le projet SpOTy.

Collecte des données

Les corpus au centre de cette présentation sont issus de terrain menés dans différents pays d'Asie du Sud-Est entre 2000 et 2024 (Vittrant 2007, Mouton 2023). Ils concernent d'une part des enregistrements de langues écrites avec des abugidas (alphasyllabaires d'origine indienne). Ces systèmes d'écriture ont parfois été tardivement standardisés (codage unicode), rarement translittérés de façon consensuelle (*cf.* birman, thaï), et posent des problèmes pour le partage des données (Seifen 2024). Sont aussi prises en compte dans notre présentation des données de langues orales non-écrites et non-décrites (*i.e.* sans grammaire écrite), des langues en danger (LED) parlées au Nord du Vietnam (hmong) (Caelen-Haumont & Vittrant 2020) ou des dialectes thaïs non-standard. Les problèmes rencontrés pour la mise à disposition des corpus sont alors d'un autre ordre : analyse grammaticale de la langue comme prérequis, choix du système de transcription, traitement de la variation par rapport à un standard (Seifen 2024).

Le tableau ci-dessous donne une idée de la (petite) taille des corpus annotés des langues étudiées actuellement disponibles dans la collection Pangloss ; certaines langues sont parfois parlées par plusieurs dizaines de millions de locuteurs comme le birman (environ 40 millions de L1).

Langues	Nombre de ressources	Nombre d'heures d'enregistrement	Nombre de documents annotés	Nombre d'heures annotées
Birman	85	31H	9	1H
Langues taïes1	205	75H	39	9H
Langues hmong2	15	2,5H	12	2H

table 1. : Les données des langues de l'étude dans la collection Pangloss

Diffusion des données

Les données orales de terrain peuvent être collectées de différentes manière (traduction de questionnaires, entretiens semi-dirigés avec ou sans outil d'élicitation, récit, conversation…) en lien avec les questions de recherche traitées. Un travail sur les sons (phonétique, phonologie) nécessite d'avoir recours, en premier lieu, à de l'élicitation de vocabulaire, tandis qu'un travail en pragmatique s'appuiera principalement sur des données collectées lors d'interactions. Dans la majorité des cas, les différentes méthodes de collecte sont cependant utilisées conjointement pour collecter des informations sur la langue et la décrire.

Car même s'il y a un engagement progressif vers la science ouverte, le travail de terrain implique un certain nombre de contraintes lors de l'élaboration de corpus annotés : des contraintes de temps qui augmentent le coût d'une donnée de terrain, mais aussi des contraintes humaines. En d'autres termes, le traitement des données (transcription, analyse, annotation grammaticales) nécessite un investissement humain important sur le long terme, mais surtout la construction d'une relation de confiance avec les locuteurs des langues étudiées.

Des archives de données de langues peu dotées sont cependant disponibles comme la collection Pangloss (Jacobson *et al* 2001, Adamou *et al* 2025). Cette archive rassemble majoritairement des enregistrements de parole spontanée, recueillis, parfois traduits et annotés par les chercheurs au fil de leurs enquêtes de terrain sur tous les continents. Nous montrerons en quoi cette archive se révèle pertinente pour la recherche en typologie.

1 La collection Pangloss regroupe des données de 9 langues taïes : khün, saek, shan, tai deng, tai don, tai lue, tai paw, tai yo, et tay khang. Une grande partie de ces données sont issues des terrains de Michel Ferlus.

2 Deux langues hmong sont représentées dans la collection Pangloss : hmong bjo et hmong noir.

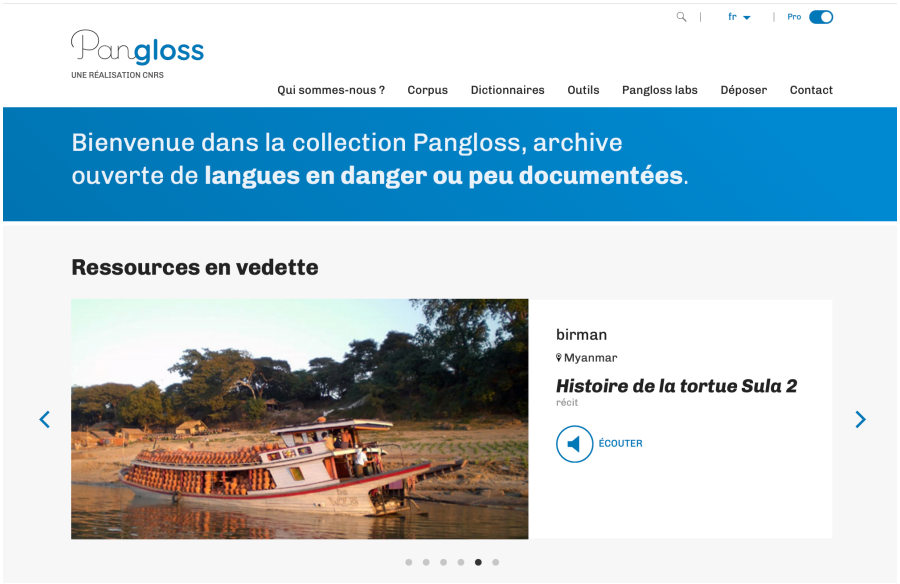


figure . 1 Interface web de la collection Pangloss

En effet, l'accès aux données de terrain va au-delà d'une mise à disposition pour la communauté scientifique et pour la communauté des locuteurs. Les données peuvent être réutilisées pour des études typologiques sur des thématiques particulières, comme avec le Projet SpOTy (Seifen *et al* 2025).

Réutilisation des données

Le projet SpOTy (*Spatial Ontology and Typology*, Labex ASLAN, Université de Lyon, 2020-2024) s'inscrit dans la lignée de deux projets en typologie du déplacement : Typologie de la Trajectoire (Fédération TUL, 2006-2008 et 2010-2011) et Deixis Dynamique (Fédération TUL, 2014-2018). Au sein de ces projets, des données de langues géographiquement, génétiquement, et typologiquement variées ont été recueillies à l'aide du matériel d'élicitation *Trajectoire* (Ishibashi *et al* 2006, Vuillermet & Kopecka 2019) par différents membres.

Le projet SpOTy a rassemblé ces données pour constituer un corpus unifié et annoté pour l'étude de l'expression du déplacement. Pour cela, les données qui étaient au départ en divers formats (corpus interlinéarisés en ligne, fichiers textes, fichiers tableurs, fichiers ELAN, fichiers Flex, voire carnets de terrain) ont été changées en un format tableur unique. Les différentes annotations sémantiques et morpho-syntaxiques (lorsqu'elles existent) ont également été normalisées, afin d'assurer l'intercomparabilité des données.

Ce projet s'inscrit dans une approche interdisciplinaire, rassemblant à la fois des linguistes et des spécialistes de l'ingénierie des connaissances. Sont utilisés au sein de ce projet de nouveau systèmes de stockage de données (espaces de stockage personnels, dits *pods*) selon les normes *Solid*. Ces *pods* permettent une diffusion plus flexible des données (non-ouvertes, ouvertes à certains utilisateurs, ouvertes à tous) et permettent l'utilisation des données dans différentes applications web (dont l'application SpOTy, portant spécifiquement sur l'expression du déplacement).

Pour finir, le projet SpOTy (et la (ré)utilisation des données) a permis de proposer une typologie fine de l'expression du déplacement dans les langues du monde (Seifen 2024), mais également des études approfondies sur les langues de cette étude, comme le birman (Vitrant *et al.* 2023) ou les langues taïes (Seifen 2024).

Références bibliographiques

Adamou, E., Guillaume, S., Michaud, A. (2025). The Pangloss Collection: Opening up research data on endangered and under-documented languages. *Language*, 101 (1), 38-59 ; DOI : http://dx.doi.org/10.1353/lan.2025.a954237.

Caelen-Haumont, G., Vittrant, A. (2020). Transcrire, écrire et formaliser en analyse phonétique, mélodique et tonale : l'exemple d'une langue d'Asie du Sud-est tonale (mo piu) et du français. *Transcrire, écrire, formaliser,* 2, 31-52. Rennes : Presses Universitaires de Rennes.

Ishibashi, M., Kopecka, A., Vuillermet, M. (2006). Trajectoire : matériel visuel pour élicitation des données linguistiques. *Fédération de Recherche en Typologie et Universaux Linguistiques.*

Jacobson, M., Michailovsky, B., Lowe, J.B. (2001). Linguistic documents synchronizing sound and text. *Speech Communication*, 33, 79-96.

Jacobson, M., Badin, F., Guillaume, S. (2015). Cocoon une plateforme pour la conservation et la diffusion de ressources orales en sciences humaines et sociales. *8es Journées Internationales de Linguistique de Corpus*. Orléans, France.

Michaud A. & Noûs C., (2024). Linguistes de terrain en quête d'éthique : de l'archivage numérique insouciant à la réflexivité permanente. *Humanités numériques* [En ligne], 10 | 2024, mis en ligne le 01 décembre 2024, consulté le 22 avril 2025. URL : http://journals.openedition.org/revuehn/4196 ; DOI : https://doi.org/10.4000/12ypu.

Mouton, L. (2023). *Esquisse grammaticale du hmong noir : étude comparative de l'expression de l'espace dans les langues hmong*, Thèse de doctorat, Sciences du Langage, Université Lumière-Lyon 2, Lyon.

Seifen, K. (2024). *Etude typologique de l'expression du déplacement : Le cas des langues taï.* Thèse de doctorat, Sciences du Langage, Université Lumière-Lyon 2, Lyon.

Seifen, K., Champin, PA., Vittrant, A. (2025). SpOTy : Une application Solid pour l'exploration de données linguistiques sur le déplacement. *Journée d'étude Approches de la typologie syntaxique à partir de corpus*. Paris, France.

Vuillermet, M., Kopecka, A. (2019). Trajectoire: A methodological tool for eliciting Path of motion. *Language Documentation & Conservation Special Publication,* 16, 97-124.

Vittrant, A. (2007). Comment constituer son corpus d'étude : exemple d'une enquête linguistique sur le birman vernaculaire. *Actes du Colloque Recueil des données en Sciences du langage et constitution de corpus : données, méthodologie, outillage,* 198-222. Nanterre, France.

Vittrant, A., Seifen, K., Champin, PA. (2023). Expression of Goal and Source in Burmese: Is there an asymmetry?. *56th International Conference on Sino-Tibetan Languages and Linguistics (ICSTLL56).* Bangkok, Thaïlande.

The Discursive Function of Pseudo-cleft Constructions from a QUD Perspective

Nariman ALIAKBAR[1], Agnès CELLE[1]
Approches Linguistiques Théoriques, Appliquées et Expérimentales : langues et cultures connectées
nariman.aliakbar@etu.u-paris.fr, agnes.celle@u-paris.fr

Introduction

This study focuses on pseudo-cleft constructions in a special type of interaction - monologic com- munication as used in TED talks. Pseudo-cleft constructions (PCs) are complex syntactic structures, typically combining a wh-clause with a focused constituent via a copula (e.g., "What John needs is a break"). While extensive research has examined their syntactic properties and structural variations (Den Dikken et al., 2006 ; Hartmann & Veenstra, 2013 ; Huddleston & Pullum, 2002), their pragmatic- discursive function in natural language use, particularly their role in managing information flow and coherence, remains significantly underexplored (Zhou & Chen, 2021). This study positions PCs not merely as syntactic artifacts but as crucial pragmatic tools speakers employ to structure discourse and guide listener interpretation.

This study investigates the discursive function of pseudo-cleft constructions in monological spoken English, focusing specifically on their interaction with the Question Under Discussion (QUD) theory (Brunetti, 2024 ; Ginzburg, 2012). The QUD theory posits that discourse is organized around implicit or explicit questions that utterances are designed to answer, thereby shaping information structure and coherence. Drawing on a corpus of 18 TED Talk transcripts (from TransQuest project), chosen for their hybrid nature between formal speech and spontaneous conversation, we explore whether PCs serve as responses to explicit or implicit questions and how their internal structure contributes to QUD dynamics.

Based on the theoretical alignment between specificational PCs and question-answer pairs (Den Dikken et al., 2006) and the principles of QUD theory, we hypothesize : (1) That specificational PCs function as direct answers to discourse questions, whether explicitly asked or implicitly understood.

(2) That the inherent focus structure of PCs, where the wh-clause often represents given information (topic/variable) and the post-copular constituent introduces new, focalized information (focus/value), actively shapes QUD salience and guides listener interpretation towards the intended query being resolved. (3) That PCs are frequently employed to resolve contextually inferred or implicit QUDs, and that listeners actively recover these unstated questions during comprehension based on linguistic and contextual cues (Westera and Rohde, 2025 ; Leonarduzzi, 2000).

To test these hypotheses, a mixed-method approach was adopted. Using regular expressions focused on "what"-cleft structures, 67 valid PCs were extracted from the corpus. A quantitative analysis mea- sured the proximity of PCs to preceding explicit questions (within a 1-6 sentence window). This was complemented by a qualitative analysis involving manual annotation to assess the semantic alignment between PCs and both explicit and potential implicit questions, categorizing alignment as Aligned, Partially Aligned, Not Aligned, or Unclear/Implied

(reflecting inferred QUDs). PCs were also cate- gorized by syntactic type (specificational vs. predicational) and referential orientation (anaphoric vs. cataphoric).

Preliminary results from the quantitative analysis indicate that while only approximately 31% of pseudo-clefts appeared within two sentences of an explicit question, suggesting that strict adjacency is not a prerequisite for a QUD relationship, a significant proportion (approximately 44%) were se- mantically aligned or partially aligned with explicit preceding questions. Crucially, an equally large proportion (approximately 44%) were found to resolve implicit or contextually inferred QUDs, provi- ding strong support for the role of listener inference. Specificational pseudo-clefts were overwhelmingly dominant (approximately 81%), aligning with theoretical predictions about their function in QUD re- solution. Qualitative case analysis further supports these findings, illustrating how PCs effectively answer questions, manage discourse coherence, simulate interaction, and strategically focus audience attention by resolving both explicit and implicit QUDs, thereby highlighting the dynamic interplay between PC structure and discourse function in monological contexts like TED Talks.

Corpus and Methodology

Corpus

The data come from 18 TED Talk transcripts (TransQuest project corpus), chosen for their hybrid nature between formal speech and spontaneous conversation. After pre-processing, 67 valid PCs were extracted using regular expressions focused on "what"-cleft structures. Each PC was manually anno- tated for proximity to prior questions (1-6 sentence window), syntactic type, referential orientation, and semantic alignment with the question.

Methodology

A mixed-method approach was adopted :

- **Quantitative Analysis :** Measuring proximity between pseudo-clefts and preceding questions.
- **Qualitative Analysis :** Assessing semantic alignment between pseudo-clefts and potential ex- plicit or implicit questions.
- **Annotation :** Manual coding of semantic alignment (Aligned, Partially Aligned, Not Aligned, Unclear/Implied).
- **Categorization :** Distinguishing specificational versus predicational pseudo-clefts, and anapho- ric versus cataphoric references.

Methodological limitations include potential regex omissions (e.g., pseudo-clefts introduced by “who” or “where”) and subjectivity in semantic alignment judgment.

Results

Feature	Count		
Total Pseudo-clefts Identified	67		
Pseudo-clefts Linked to Questions	43		
Pseudo-clefts Within 1–2 Sentences of a Question	21	(≈31%)	Specificational
Pseudo-clefts	54		Predicational
Pseudo-clefts	13		

table 1. : Quantitative Results

Only 31% of pseudo-clefts appeared within two sentences of an explicit question, suggesting a significant role for implicit QUDs.

Alignment Category	Number of Cases
Aligned	24
Partially Aligned	6
Unclear/Implied	30
Not Aligned	8

table 2. : Semantic Alignment

Argument

Quantitative analysis reveals that 31% of PCs occurred within 1-2 sentences of an explicit question. However, semantic coding showed that 44% were aligned, and 44% resolved inferred QUDs. Specifi- cational PCs were overwhelmingly dominant (80.6%), and most were anaphoric, referencing earlier discourse. This reflects the form-function alignment of PCs in QUD resolution.

The qualitative analysis of the data further illustrates how pseudo-cleft constructions operate at the interface of syntax, discourse, and pragmatics by aligning with, reframing, or expanding ongoing QUDs. For instance, in (1) *What did Darwin say ? – What Darwin said was something like this : if you have creatures. . .*, the pseudo-cleft occurs immediately after the question as a direct answer. Note that the direct answer comes after the colon. The pseudo-cleft construction (wh clause + postcopular constituent) rather functions as a resumptive structure that introduces an answer. Although predica- tional in type, its discourse function is unmistakably specificational : the wh-clause ("what Darwin said") retrieves given information from the question, while the post-copular clause provides the focal value. This confirms Hypothesis 1 by showing how PCs can be used as straightforward responses to explicit QUDs (cf. Declerck, 1988 ; Den Dikken et al., 2006). Declerck (1988) did not argue that PCs inherently respond to questions ; rather, he analyzed them as two-part constructions, with the wh- clause functioning like a question and the post-copular constituent providing its answer. Our analysis extends this by showing how in discourse such a structure can map directly onto explicit or implicit QUDs.

In other cases, PCs do not supply a direct answer but instead redirect the discourse by foregroun- ding a related element. In (2) *What do we know ? – . . .but what is remarkable is how easy it is to make this impatience go away...*, the PC does not directly answer the epistemic question "what do we know." Instead, it reframes the QUD around what is "remarkable", shifting the focal point away from knowledge toward evaluation. This type of partial alignment illustrates how the inherent focus struc- ture of PCs can guide salience and reorient the listener toward an alternative interpretation, thereby supporting Hypothesis 2. Such reframing strategies resonate with observations in Prince (1978) on the discourse-organizing role of pseudo-clefts and their importance in the listener's recovery of implicit questions, potentially filling an informational void (Leonarduzzi, 2000).

A further pattern involves cases where PCs both respond to the explicit QUD and simultaneously reframe it, thereby addressing implicit rhetorical dimensions. The following case illustrates this dual role.

Example (3) *How do we know what to do if we don't have a moral framework ? – What we really need is a moral operating system...* shows a specificational PC that functions both as an explicit answer and as a normative reframing of the question. Rather than addressing the epistemic difficulty posed in the question, the construction asserts the necessity of a "moral

operating system." In doing so, it simultaneously resolves the explicit QUD and evokes an implicit one concerning the conditions for moral action. In other words, the question implies that "we don't know what to do without a moral framework." Here the PC indirectly answers this rhetorical question, implying that the direct answer is 'no', aligning semantically with the rhetorical questions. Even rhetorical questions can function as QUD-triggers (cf. Celle, 2009 ; Westera and Rohde, 2019). This pattern aligns with Hypothesis 2, where the PC's focus structure shapes QUD salience, and with Hypothesis 3, which emphasizes the inferential role of the listener in reconstructing unstated questions (cf. Berthe and Gaudy-Campbell, 2024).

While our annotation scheme was tailored to the data, it should be further refined in future work by aligning with established frameworks, such as Westera and Rohde's (2019, 2025) model of evoked QUDs or Ginzburg et al.'s (2019) response space model. A mixed model may also better capture the interplay between explicit and inferred QUDs.

This study contributes to the broader understanding of how monological discourse incorporates dialogical strategies. According to our data, PCs manage coherence, simulate question-answer ex- changes, and foreground focal content, thereby enhancing audience orientation. From a theoretical standpoint, this research enriches QUD-based models of information structure and demonstrates the explanatory power of integrating constructional approaches with pragmatic-discursive analysis.

References

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, 24(1), 65–87. https ://doi.org/10.1017/S0022226700011567

Berthe, F., & Gaudy-Campbell, I. (2024). Au-delà de la focalisation : La pseudo-clivée comme stratégie de recherche d'adhésion. *Anglophonia. French Journal of English Linguistics*. https ://doi.org/10.4000/12poa

Brunetti, L. (2024). Contrast in a QUD-based information-structure model. In J. Brysbaert & K. Lahousse (Eds.), *On the Role of Contrast in Information Structure* (pp. 191–224). Berlin : De Gruyter. https ://doi.org/10.1515/9783110986594- 008

Brunetti, L. (2024). *Information structure in Romance : From sentence to discourse*. Paris : Université Paris Cité.

Celle, A. (2009). Question, mise en question : La traduction de l'interrogation dans le discours théorique. *Revue française de linguistique appliquée*, XIV(1), 39–52. https ://doi.org/10.3917/rfla.141.0039

Declerck, R. (1988). *Studies on copular sentences, clefts and pseudo-clefts*. Leuven : Leuven University Press.

Den Dikken, M., Everaert, M., & Riemsdijk, H. C. (2006). Pseudoclefts and Other Specificational Copular Sentences. In *The Wiley Blackwell Companion to Syntax, Second Edition* (pp. 292–409). Malden, MA : Blackwell Publishing. https ://doi.org/10.1002/9781118358733.wbsyncom001

Ginzburg, J. (2012). *The interactive stance : Meaning for conversation*. Oxford : Oxford University Press.

Ginzburg, J., Yusupujiang, Z., Li, C., Ren, K., & Łupkowski, P. (2019). Characterizing the Response Space of Questions : A Corpus Study for English and Polish. *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 320–330. https ://doi.org/10.18653/v1/W19-5937

Goldberg, A. E. (2007). *Constructions : A construction grammar approach to argument structure* (5th ed.). Chicago : The University of Chicago Press.

Goldberg, A. E. (2014). Goldberg's Construction Grammar. In K. Ramonda (Ed.), *The Bloomsbury Companion to Cognitive Linguistics* (pp. 60–71). London : Bloomsbury Publishing. https ://doi.org/10.5040/9781472593689.ch- 004

Gundel, J. K. (1977). Where do Cleft Sentences Come from ? *Language*, 53(3), 543. https ://doi.org/10.2307/413176 Hartmann, K., & Veenstra, T. (2013). *Cleft structures*. Amsterdam : John Benjamins Publishing Company.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language* (1st ed.). Cambridge : Cambridge University Press. https ://doi.org/10.1017/9781316423530

Leonarduzzi, L. (2000a). *La subordonnée interrogative en anglais contemporain*. Aix-Marseille : Université de Provence.

Leonarduzzi, L. (2000b). Interrogatives in contemporary English. *Proceedings of the 5th International Conference on English Language and Literature*. Aix-en-Provence : Université de Provence.

Prince, E. F. (1978). A Comparison of Wh-Clefts and it-Clefts in Discourse. *Language*, 54(4), 883. https ://doi.org/10.2307/413238

Westera, M., Mayol, L., & Rohde, H. (2020). TED-Q : TED Talks and the Questions they Evoke. *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France.

Westera, M., & Rohde, H. (2019). Asking between the lines : Elicitation of evoked questions in text. *Proceedings of the Amsterdam Colloquium*, 397–406. Amsterdam, The Netherlands.

Westera, M., & Rohde, H. (2025). Asking between the lines : Elicitation of evoked questions in text. *Journal of Pragmatics*, 210, 112–134.

Zhou, H., & Chen, M. (2021). What Still Needs to be Noted : Pseudo-Clefts in the Academic Discourse of Applied Linguistics. *Frontiers in Psychology*, 12, 672349. https ://doi.org/10.3389/fpsyg.2021.672349

Towards extracting patterns of modal collocations to syntactic complexities for assessing truth-conditional alignment

Tao MA
Shanghai Sanda University
taoma@sandau.edu.cn

This study examines the distributional patterns of modal verb-adverb collocations in political contingency discourse by using a specialized 12-million-token corpus of political holding statements, and compares these patterns against a 1-billion-token baseline corpus to measure register-specific deviations from default states of agreement between lexical bundles and syntactic structures. It is hypothesized that truth conditions are norm-referenced and governed by register-specific agreement between semantic roles and syntactic features. Analyzing syntactic patterns in modal collocations reveals the alignment between truth conditions and the formal representation of modality in register-specific discourse. Notably, the elevated syntactic complexity of asymmetrical pairings like *should-likely, would-certainly, would-actually* and *should-probably,* points to a compensatory effect within these collocations. In political contingency discourse, where truth conditions are paramount, this supports the view that modal operators are pragmatically instantiated through norm-referenced register-specific alignments.

Transfert séquentiel avec BERT au service de l'annotation sémantique : une approche critique de la linguistique augmentée

Guillaume Desagulier [1], Redinela Biba [1,2]
[1] Laboratoire CLIMAS, Université Bordeaux Montaigne
[2] Université de Tirana
guillaume.desagulier@u-bordeaux-montaigne.fr

Introduction

Selon Do, Ollion et Shen (2022), les avancées récentes en traitement automatique des langues offrent aux chercheurs en sciences sociales la possibilité d'annoter automatiquement et avec précision "des millions de textes". Cette capacité repose sur l'apprentissage par transfert séquentiel, une famille de modèles d'apprentissage profond dans le sillage de BERT (Devlin et al. 2019), qui s'appuient sur l'architecture Transformer (Vaswani et al. 2017) et ses mécanismes d'attention pour capturer les dépendances à longue distance. Ces modèles sont pré-entraînés sur d'immenses corpus non annotés, puis adaptés à des tâches spécifiques avec peu de données annotées. Do, Ollion et Shen démontrent qu'ils atteignent désormais une précision comparable à celle des codeurs humains, sans effet de fatigue, et peuvent être personnalisés pour les besoins spécifiques des sciences sociales.

En sémantique de corpus, les études fondées sur de très grands corpus produisent souvent d'imposants jeux de données, qu'il est alors peu productif, voire impossible, d'annoter manuellement dans leur intégralité. Il est donc particulièrement utile de recourir à des classifieurs automatiques capables d'opérer à grande échelle. Si l'exploitation de vastes quantités de données présente un réel intérêt, elle ne doit cependant pas se faire au détriment de la précision de l'annotation, ni, par extension, de l'analyse linguistique. Les promesses des grands modèles de langage méritent donc d'être examinées sous un angle critique.

Notre étude s'attache à évaluer si l'apprentissage par transfert séquentiel, vanté par Do, Ollion et Shen (2022), représente une véritable révolution méthodologique. Si la qualité de l'annotation proposée par cette méthode est satisfaisante, elle facilitera le recours à des jeux de données issus de très grands corpus.

Pour illustrer ces enjeux, nous proposons une étude de cas portant sur l'adjectif *cool* en anglais américain. Le choix de ce lexème se justifie par sa très grande polysémie et son évolution diachronique marquée, qui en font un terrain d'observation privilégié pour évaluer la capacité des modèles automatiques à saisir les nuances sémantiques fines. Notre extraction compte plus de 20 000 occurrences de *cool*. De par sa taille, l'annotation sémantique de la totalité du jeu de données est particulièrement ardue. Seul un échantillon représentant un vingtième du corpus

total a donc fait l'objet d'une double annotation manuelle. Le reste du corpus a été traité automatiquement à l'aide d'un algorithme s'appuyant sur XML-RoBERTa[1].

Si les performances du modèle s'avèrent globalement satisfaisantes, l'annotation automatique reproduit certaines difficultés observées chez les annotateurs humains, notamment au niveau de la délimitation entre les sens non-littéraux et émotionnels de *cool*. Cette convergence dans les zones d'incertitude suggère que les défis interprétatifs inhérents à la polysémie ne sont pas entièrement résolus par l'automatisation, mais plutôt déplacés du niveau de l'annotation manuelle vers celui de l'apprentissage automatique.

Le reste de cet article est organisé comme suit : nous présenterons d'abord notre corpus d'étude et notre méthodologie d'annotation, puis nous exposerons nos résultats et leur discussion, avant de conclure sur les implications de cette recherche pour l'avenir de l'annotation automatique en sémantique de corpus.

Étude de cas, corpus et méthodologie

Notre étude s'appuie sur les fichiers sources du Corpus of Historical American English (Davies 2010), qui comprend plus de 475 millions de mots de textes en anglais américains sur une période allant des décennies 1820 à 2010. Nous avons limité la requête aux décennies 1900-2010, soit 345 777 090 mots.

Un script d'extraction rédigé en R nous a permis d'extraire 20 157 occurrences du lexème adjectival *cool* en contexte (fenêtre de ± 10 mots). Les variables relevant des périodes historiques et des genres textuels ont été intégrées de manière à bénéficier d'une base multifactorielle pour analyser l'évolution sémantique de ce lexème.

Un échantillon représentatif de 1 008 occurrences de l'adjectif *cool* a été annoté manuellement selon trois catégories sémantiques :

- *Basic* : sens littéral lié à la température
  1. (...) the limb should be kept **cool** to reduce its oxygen requirements to a minimum.
- *Nonliteral* : usages métaphoriques ou figurés
  2. (...) you got my word he's **cool**. He's been on nine jobs with me.
- *Emotion* : sens exprimant une attitude ou un état émotionnel
  3. (...) Giulio Andreotti, 57, a **cool** and analytical Roman (...)

Afin d'évaluer la cohérence et la reproductibilité du système d'annotation, un second annotateur a procédé indépendamment à l'étiquetage d'un sous-ensemble aléatoire et équilibré de 229 tokens, représentant 22% de l'échantillon et répartis de manière homogène entre genres et décennies. L'accord inter-annotateurs a été mesuré au moyen du κ de Cohen. Le score obtenu (0,71) correspond à un accord substantiel d'après Landis et Koch (1977). Le système d'annotation se révèle donc fiable et intelligible pour les annotateurs.

Notre approche repose sur l'entraînement d'un modèle de classification automatique pour annoter les occurrences de l'adjectif anglais *cool* selon les trois catégories sémantiques présentées ci-dessus. Nous avons fait le choix d'utiliser le modèle pré-entraîné XLM-RoBERTa base, un réseau de neurones transformeur développé par Facebook AI, reconnu pour ses

1 https://huggingface.co/docs/transformers/model_doc/xlm-roberta

performances sur les tâches de classification de séquences multilingues. L'entraînement se fait sur la base des 1 008 occurrences de *cool* annotées manuellement. Une fois entraîné, le modèle est utilisé pour annoter automatiquement les 19 149 occurrences restantes.

La méthodologie comprend plusieurs étapes d'optimisation technique : division stratifiée des données (85% entraînement, 15% test) qui préserve la distribution des catégories, tokenisation avec troncature à 256 unités lexicales (découpage du texte en éléments analysables par le modèle), et entraînement avec arrêt précoce (2 époques) pour éviter le surapprentissage. L'évaluation repose sur des métriques statistiques standard : la précision (pourcentage de classifications correctes) et le F1-score pondéré (moyenne harmonique entre précision et rappel), avec un objectif de F1 $\geqslant$ 0,80. La validation qualitative s'appuie sur l'analyse des matrices de confusion qui visualisent les erreurs de classification, ce qui permet d'identifier les confusions sémantiques entre catégories.

Résultats et discussions

L'entraînement de notre modèle fait apparaître une amélioration constante des performances sur 5 époques (Tableau 1). Les résultats indiquent une bonne convergence avec un pic de performance à l'époque 4, suivi d'un léger surapprentissage à l'époque 5. L'apprentissage s'est donc interrompu après quatre époques.

Époque	Perte d'entraînement	Perte de validation	Précision	F1-score (macro)
1	0,94	0,70	0,73	0,72
2	0,66	0,62	0,74	0,74
3	0,51	0,58	0,79	0,79
4	0,33	0,51	0,84	0,84
5	0,32	0,52	0,82	0,82

table 1. : Évolution des métriques d'apprentissage sur 5 époques avec XLM-RoBERTa

L'analyse détaillée des performances par classe révèle des résultats contrastés. La catégorie Basic obtient d'excellents résultats avec une précision et un rappel de 0,93, ce qui reflète la nature relativement moins ambiguë des usages littéraux liés à la température. La catégorie *Emotion* présente une précision correcte de 0,83 mais un rappel plus faible de 0,75, signe que le modèle a tendance à être conservateur dans la détection de ces expressions. La catégorie *Nonliteral* montre les performances les plus modestes avec une précision de 0,7 et un rappel de 0,76, ce qui se comprend aisément de par la complexité inhérente à la détection des usages figurés et métaphoriques.

L'application du modèle aux données non annotées révèle une distribution intéressante des prédictions avec une confiance moyenne élevée de 0,96. La répartition des prédictions montre une prédominance des usages *Basic* (46,7%), suivis des usages *Nonliteral* (36%) et *Emotion* (17,3%). Cette distribution suggère que les usages littéraux de la température demeurent, sans surprise, majoritaires dans notre corpus, tout en révélant une proportion substantielle d'expressions figurées.

Les matrices de confusion ainsi que le score F1 montrent un degré d'accord relativement élevé entre les catégories sémantiques du modèle et des annotateurs humains (Figure 1). Pour les usages basiques, le modèle s'accorde fortement avec les deux annotateurs. Pour les catégories *Emotion* et *Nonliteral*, qui sont les plus difficiles à distinguer, le modèle s'aligne davantage avec Annot1 qu'avec Annot2.

Concernant à présent la performance du modèle sur les données non annotées, l'analyse met en évidence trois points de vigilance principaux. Premièrement, le modèle confond parfois les

expressions figurées contenant des références à la température avec des usages littéraux car il a tendance à privilégier les associations lexicales fréquentes au détriment du sens contextuel. Deuxièmement, un chevauchement conceptuel entre les catégories *Emotion* et *Nonliteral* entraîne des confusions pour des expressions polysémiques comme *a cool attitude* ou *stay cool*. Ceci n'est pas un défaut en soi car même les annotateurs humains peuvent éprouver des difficultés face aux frontières floues entre domaines sémantiques. Enfin, le modèle présente une tendance à la surclassification dans la catégorie *Basic* même lorsque le contexte suggère un usage métaphorique.

Les limites du modèle peuvent s'expliquer par le fait que XLM-RoBERTa a été entraîné sur des phrases complètes, alors que nous n'utilisons que des extraits de phrases. Cela qui peut entraver l'accès au contexte sémantique nécessaire pour désambiguïser les usages polysémiques. De plus, la nature déséquilibrée de notre jeu de données d'entraînement, avec une surreprésentation de l'étiquette *Basic*, a certainement biaisé les prédictions du modèle vers cette catégorie majoritaire.

Une ACM réalisée sur l'ensemble du jeu de données annoté automatiquement, enrichi des variables contextuelles extraites du corpus (année de publication, genre textuel) et des contextes syntaxiques (attribut/épithète), révèle des tendances cohérentes (Figure 2). L'adjectif *cool* fait l'objet d'une transformation sémantique progressive. Cette trajectoire indique un processus de grammaticalisation par lequel *cool* transite d'un adjectif descriptif vers un marqueur expressif polyvalent, dont la spécialisation s'ajuste aux contraintes communicatives de chaque genre textuel.

Conclusion

L'apprentissage par transfert séquentiel offre des possibilités prometteuses pour l'annotation sémantique à grande échelle en linguistique de corpus. Toutefois, bien que ses performances soient exploitables, l'annotation automatique présente des limites qualitatives qui, sans toutefois compromettre la pertinence globale du modèle, nécessitent une supervision humaine continue.

L'approche la plus efficace demeure donc une méthode hybride : annotation manuelle rigoureuse, annotation automatique par apprentissage profond, vérification des résultats selon la méthode d'échantillonnage proposée, puis perfectionnement itératif du modèle. Cette linguistique augmentée n'est pas prête à remplacer l'expertise humaine. Elle la complète et permet malgré tout aux linguistes d'explorer des volumes de données textuelles auparavant inaccessibles tout en préservant la rigueur analytique nécessaire.

Références bibliographiques

Davies, M. (2010). *The Corpus of Historical American English (COHA): 475+ million words, 1820s-2019*. https://www.english-corpora.org/coha/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Do, S., Ollion, É., & Shen, R. (2024). The augmented social scientist: Using sequential transfer learning to annotate millions of texts with human-level accuracy. *Sociological Methods & Research*, *53*(3), 1167-1200.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Figures

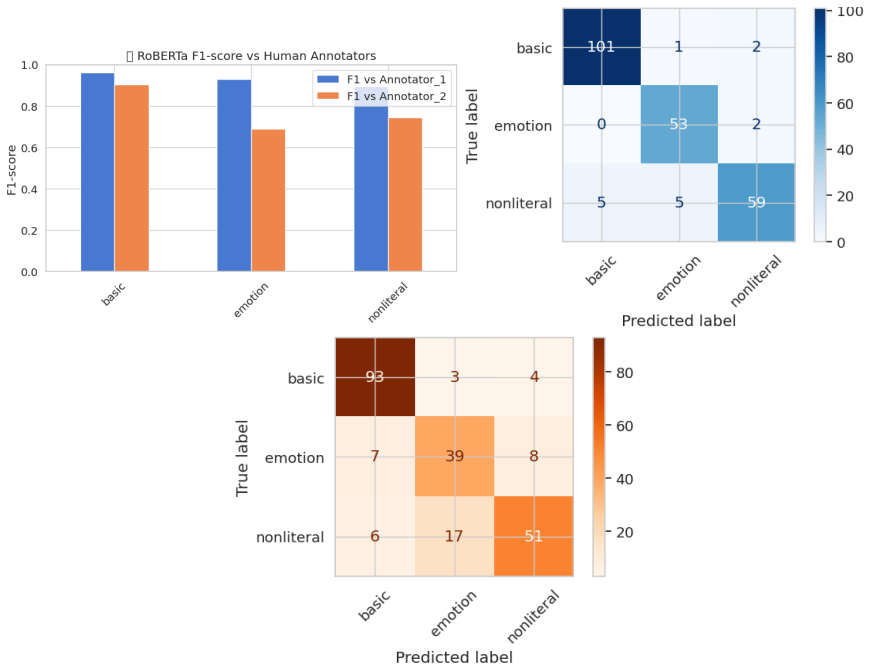


figure . 1 Comparaison des performances entre XML-RoBERTa et les annotateurs

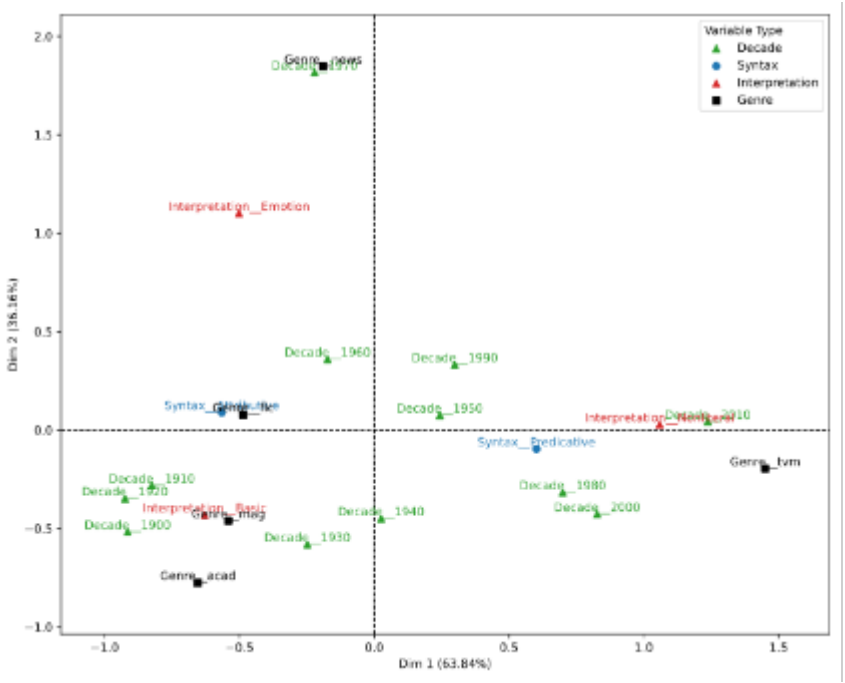


figure . 2 ACM : exploration finale des sens de *cool* à travers les décennies et les genres textuels du COHA

Tri des concordances : comparaison des critères lexicosyntaxiques et des plongements lexicaux

Olivier Kraif [1]
[1] Laboratoire LIDILEM, Université Grenoble Alpes
olivier.kraif@univ-grenobles-alpes.fr

Introduction

Dans l'exploration de corpus textuels, le tri des concordances en fonction des contextes est une fonctionnalité standard de la plupart des concordanciers (Pincemin et al., 2006). Par exemple, il est généralement possible de regrouper les concordances autour d'un pivot verbal en fonction des prépositions ou conjonctions qui suivent immédiatement le verbe (tri par le contexte droit), ou bien des pronoms ou adverbes qui le précèdent (tri par le contexte gauche). Si certains concordanciers comme Antconc (Anthony, 2013) proposent des options plus avancées, tels que s'appuyer sur les formes en position -1, -2, -3 à gauche ou +1, +2, +3 à droite, par rapport au pivot, il faut admettre que la caractérisation des contextes reste relativement pauvre.

Pourtant, regrouper les contextes de manière plus fine semble indispensable si on veut pouvoir interpréter finement les pivots en incluant toutes les dimensions de ce que Sinclair (1991) appelle les "extended units of meaning" : les collocations lexicales (associations privilégiées avec d'autres lexèmes), mais aussi les colligations (associations privilégiées avec des positions syntaxiques), les classes sémantiques (associations privilégiées avec des traits sémantiques) et la prosodie sémantique (connotations, polarité, ironie, etc., cf. Louw, 1993).

Le classement lexicosyntaxique des contextes est même au coeur de la méthode de description promue par Hanks (2004) sous le nom de CPA, pour *Corpus Pattern Analysis*. Dans cette méthode, l'interprétation d'un lexème doit être guidée par un ensemble de paramètres lexico-syntaxiques caractérisant le contexte immédiat de l'unité. Typiquement, pour un verbe, on cherchera à décrire non seulement sa structure argumentale, mais aussi les indices lexicosyntaxiques liés à ses arguments : « Verb patterns consist not only of the basic 'argument structure' or 'valency structure' of each verb (typically with semantic values stated for each of the elements), but also of subvalency features, where relevant, such as the presence or absence of a determiner in noun phrases constituting a direct object » (Hanks, 2004:87).

La description des patterns nominaux est relativement simple, car elle incorpore surtout la dimensions collocationnelle : « Noun patterns consist of a number of corpus-derived gnomic statements, into which the most significant collocates are grouped and incorporated. » (Hanks, ibid.). Les patterns verbaux, quant à eux présentent des configurations syntaxiquement plus complexes, faisant intervenir le sujet, la structure argumentale, les éventuels adverbes, et les classes sémantiques des arguments.

Ainsi, un *pattern* verbal pourra être défini non seulement par le type de sujet, la présence ou l'absence d'un objet, éventuellement précédé d'une préposition, mais aussi par les traits sémantiques portés par les éventuels objets.

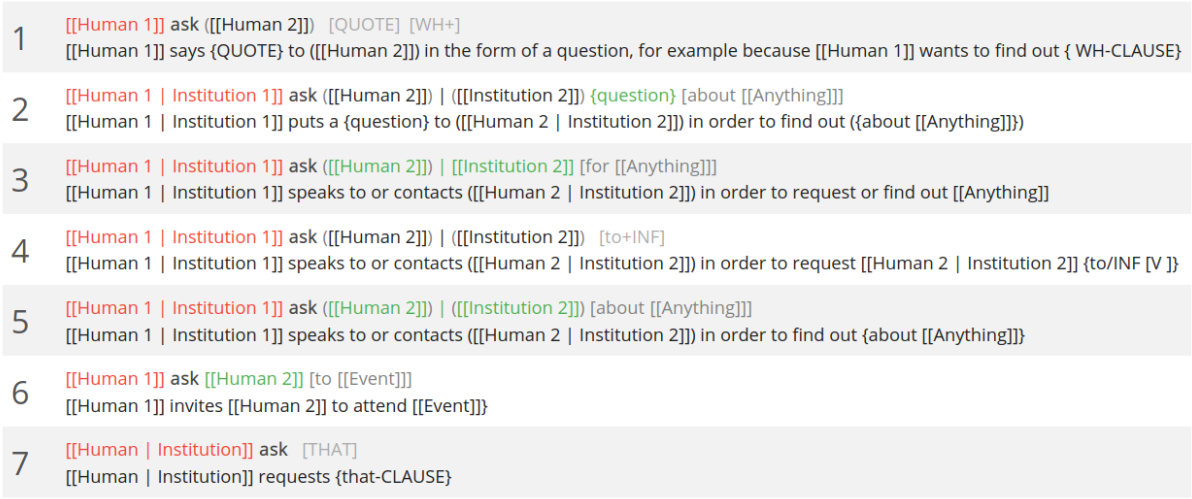


figure . 1 Exemple de description CPA pour le verbe ask

La figure 1 donne un extrait de la description du verbe *ask* issue du Pattern Dictionary of English Verbs (Hanks, 2013). Les étiquettes sémantiques définissant les classes lexicales s'inspirent du modèle du lexique génératif de Pustejovsky (1995).

La méthode décrite par Hanks est totalement empirique et analogique : on commence par extraire une concordance comportant suffisamment d'occurrences tirées au hasard (typiquement entre 200 et 1000) pour décrire un verbe. Puis on regroupe les occurrences en fonction de leur affinité lexicosyntaxique, et de l'interprétation du pivot en contexte. De la sorte, on fait naturellement émerger des *patterns* qui correspondent aux différentes acceptions du pivot.

Cette proposition de communication vise à répondre à la question suivante : quelle méthode appliquer pour effectuer un tri automatique des concordances qui permettrait de regrouper les occurrences en fonction de ces variables contextuelles, en utilisant toute la richesse des paramètres syntaxiques, mais aussi sémantiques, à disposition ?

Méthodologie

Pour répondre à cette question, nous proposons de comparer deux approches : une approche symbolique, s'appuyant sur l'analyse syntaxique en dépendances des phrases constituant la concordance, et une approche basée sur un modèle de langage préentrainé, permettant de calculer des plongements contextuels des pivots.

Pour cette expérimentation, nous avons intégré ces deux fonctionnalités dans la plate-forme du Lexicoscope 2.0 (Kraif, 2019), qui permet notamment d'explorer des corpus analysés en dépendances en s'appuyant sur un langage de requête enrichi, le langage TQL (Abouwarda & Kraif, 2024).

Nous avons choisi d'effectuer des tests sur la version analysée avec Stanza (Qi et al., 2020), du corpus Phraseorom (Diwersy et al., 2021), qui présente l'intérêt d'être génériquement homogène (corpus romanesque) et de taille suffisante (107M de tokens) pour présenter une combinatoire riche des contextes.

Groupement lexicosyntaxique des contextes

Pour le groupement lexicosyntaxique des contextes, nous avons développé une mesure de similarité syntaxique dans une perspective assez proche de Wang et al. (2017) : on examine l'ensemble des relations de dépendances partant du pivot, qu'il soit en position de gouverneur ou de dépendant syntaxique. Chaque relation est typée par son étiquette et par le lemme pointé par la relation.

Nous testons ainsi différentes conceptions de la similarité :

- soit par abstraction du contexte, en ignorant le lemme pointé par la relation pour certaines catégories (les mots pleins, essentiellement les noms, noms propres, verbes et adjectifs, sont alors remplacés par leurs étiquettes grammaticales). Dans ce cas, les groupes sont simplement défini par l'identité du contexte une fois abstrait.
- soit par calcul d'une mesure de similarité s'appuyant sur le calcul normalisé du nombre de relations communes (suivant une mesure de Dice). Dans ce cas de figure, nous proposons optionnellement de s'appuyer sur un calcul de similarité lexicale pour les lemmes, en nous basant sur les vecteurs dérivés de FastText (Bojanowski et al. 2017). Nous appliquons alors un algorithme de partitionnement hiérarchique, en fixant un seuil de similarité pour définir les clusters.

Calcul des plongements par transformers

Une autre voie consiste à calculer le plongement contextuel du pivot, pour chaque exemple de concordance, via un *transformer* de type BERT (Devlin et al., 2017). Un tel plongement condense, sous la forme d'un vecteur numérique, à la fois l'interprétation sémantique du pivot en contexte, mais aussi, via les matrices d'attention qui ont servi à la construction de ce plongement, les propriétés du contexte pertinentes pour l'élaboration de cette interprétation. A partir des vecteurs obtenus pour chaque exemple de concordance, il est possible de mettre en oeuvre un algorithme de partitionnement hiérarchique ascendant afin d'obtenir des groupes pour les vecteurs les plus similaires (selon la mesure du cosinus). Pour cette expérimentation, nous avons utilisé le modèle Flaubert Large Cased (Le et al., 2020).

Résultats

Une analyse qualitative des résultats indique que la méthode de regroupements des contextes identiques, une fois abstrait des lemmes pleins, apparait comme la moins couteuse en calcul et la plus efficace pour obtenir une réponse rapide de tri des concordances dans une consultation en ligne. Les clusters obtenus sont toutefois assez petits, et nécessitent d'être regroupés dans un second temps si l'on veut mettre en oeuvre une description de type CPA. Le calcul de similarité s'appuyant sur Dice permet d'obtenir des groupement plus larges, mais aussi moins cohérents.

Quant à l'utilisation des plongements, elle fournit des groupements moins facilement interprétables. La méthode semble néanmoins prometteuse si on applique une requête pour un pivot susceptible de varier (p. ex. une requête du type "ne cache pas son" + NOM, qui renvoie des noms d'affects très variés, tels que *admiration, joie, colère, tristesse, exaspération, impatience*, etc.), car elle permet alors de regrouper les variantes assez finement en fonction de leur similarité sémantique. Des exemples concrets des ces observations seront fournis lors de la présentation.

Références bibliographiques

Abouwarda, R., Kraif, O. (2024). Utilisation de requêtes syntaxiques pour la terminologie : une étude de cas dans les domaines de la psychologie et de la psychiatrie. *9e Congrès Mondial de Linguistique Française*, SHS Web Conf. Volume 191, 2024.

Anthony, L. (2013). Developing AntConc for a new generation of corpus linguists. *Proceedings of the Corpus Linguistics Conference (CL 2013)*, 14–16.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching Word Vectors with Subword Information. In *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 201.

Devlin J., Chang M.-W., Lee K ., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pages 4171–4186.

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, 1–22.

Kraif, O. (2019). Explorer la combinatoire lexico-syntaxique des mots et expressions avec le Lexicoscope. In Max Silberstein (dir.), *Langue française*, N° 203, Armand Colin, p. 67-82.

Hanks, P. (2004). Corpus pattern analysis. In *Euralex Proceedings*, volume 1, pages 87–98.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France.

Louw, B. (1993). Irony in the Text or Insincerity in the Writer?—The Diagnostic Potential of Semantic Prosodies. In M. Baker, G. Francis, & E. Togni-ni-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 157-176). Amsterdam: Benjamins. https://doi.org/10.1075/z.64.11lou

Pincemin, B., Issac, F., Chanove, M., Mathieu-Colasin, M. (2006). Concordanciers : Thème et variations. *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2006)*, Jean-Marie VIPREY et al. (éds), Besançon : Presses Universitaires de Franche-Comté, ISBN 2.84867130.0, vol. II, pp. 773-784.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J, Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*. https://arxiv.org/abs/2003.07082

Sinclair, J (1991). *Corpus, concordance, collocation*. Oxford University Press

Wang, I., Kahane, S., Tellier, I. (2016). From built examples to attested examples: a syntax-based query system for non-specialists. *PACLIC30*, Jong-Bok Kim, Oct 2016, Seoul, South Korea.

TuLHis1 : présentation d'un corpus de lectures partagées et premiers résultats

Céline Dugua[1] Eva Troupillon[1], Flora Badin[1], Olivier Baude[2]
[1] Laboratoire Ligérien de Linguistique (UMR7270), Université d'Orléans
[2]IR* Huma-Num
celine.dugua@univ-orleans.fr, eva.troupillon@etu.univ-orleans.fr, flora.badin@univ-orleans.fr, Olivier.baude@huma-num.fr

Introduction

La lecture partagée correspond à des scènes lors desquelles un adulte lit une histoire à un enfant non-lecteur (Frier, 2006). Cette question est largement étudiée du point de vue de son intérêt à la fois pour le développement langagier de l'enfant mais aussi pour son entrée dans la littératie (Grossmann, 2006).

L'angle particulier que nous explorons dans le corpus TuLHis1 (*« Tu me lis une histoire ! »*) consiste à rendre compte de la façon dont l'adulte oralise cette lecture face à un ou des enfants, en élargissant la définition aux enfants lecteurs. Observer finement les façons de pratiquer la lecture partagée présente un intérêt majeur dans le cadre des études sur la parole adressée à l'enfant (ou *Child directed speech*, CDS). Aujourd'hui, on connait très bien les caractéristiques du CDS, tant du point de vue prosodique, lexical, syntaxique (Snow, 1977, 1999 ; Cameron-Faulkner & Hickey, 2011). On connait l'importance de cette parole pour son développement langagier, et son caractère évolutif avec l'âge de l'enfant (Foulkes et al. 2005). Or, les lectures partagées constituent une forme de CDS mais dont les caractéristiques linguistiques demeurent peu connues. Une particularité de la lecture partagée réside dans la présence d'un support écrit et illustré entre le lecteur et l'enfant. Nous nous intéresserons ici à l'influence que peut avoir la présence de cette forme écrite entre l'adulte et l'enfant sur le langage adressé à ce dernier. Evidemment, elle contraint l'échange autour du texte, puisque les lecteurs vont avant tout oraliser le texte écrit ; toutefois, on remarque qu'ils s'approprient le texte, qu'ils peuvent s'en détacher en partie, s'en servir de prétexte à d'autres échanges, etc. Autant de spécificités que le corpus TuLHis1 permettra de décrire.

L'oralisation de lectures partagées sera étudiée à travers un phénomène phonologique particulier : l'usage des liaisons. Les liaisons se caractérisent par la production d'une consonne (dite consonne de liaison) à la frontière de deux mots ; par exemple, une liaison /n/ apparait entre « un » et « avion » dans /œnavjɔ̃/, une liaison en /z/ entre « les » et « avis » dans /lezavi/, une liaison en /t/ entre « tout » et « à » dans /tutaku/. Selon le contexte syntaxique, les liaisons peuvent être systématiques (liaisons également appelées obligatoires) ou variables (liaisons facultatives). Ces dernières sont connues pour être des variables sociolinguistiques dont l'usage varie selon les caractéristiques des locuteurs et les types de situations notamment (Gadet, 1989). Le fonctionnement des liaisons en français s'inscrit dans l'histoire de la langue, de sa prononciation et de la distance que celle-ci a pris avec l'orthographe au fil du temps (Langlard, 1928). Observer les liaisons à travers la lecture permet de réintégrer cette caractéristique graphique dans l'étude de la liaison.

Dans cette étude, nous observerons l'ensemble des liaisons en croisant leur usage avec des facteurs sociolinguistiques caractéristiques du lecteur, de l'enfant et également avec des

facteurs liés à l'album. Nous observerons aussi les cas de liaisons « inattendues » ou pataquès (Coutanson, 2023), c'est-à-dire des liaisons qui ne correspondent pas aux caractéristiques des deux mots impliqués, par exemple un ajout de consonne de liaison entre « si » et « important » dans /sizɛ̃pɔʁtɑ̃/ ou une substitution de consonne entre « vingt » et « enfants » dans /vɛ̃zɑ̃fɑ̃/. On peut faire l'hypothèse qu'en lecture, les liaisons inattendues peuvent être plus fréquentes qu'en parole spontanée en raison d'une plus grande surveillance du lecteur.

Corpus et méthodologie

Corpus

Le corpus TuLHis1 a pour ambition de collecter et analyser des lectures partagées en cherchant à diversifier les lecteurs et les enfants et en contrôlant les albums lus, à partir de la constitution d'une bibliothèque d'une cinquantaine d'albums, à la disposition des participants. Cette première sélection d'albums va être amenée à évoluer en partie afin de répondre plus largement aux attentes des enfants selon leurs âges. Dans cette bibliothèque, un album tient une place particulière « Histoires de Lou » de Natali Fortier publié aux éditions du Rouergue. Sa création littéraire a été l'occasion d'une collaboration avec l'auteure-illustratrice. Cet album est constitué de 12 petites histoires qui peuvent être lues indépendamment les unes des autres. Nous encouragions les participants à lire au moins une histoire de leur choix de cet album.

Au jour de la rédaction de cette soumission, le corpus collecté est en cours de traitement et représente environ 6 heures d'enregistrements. Il est constitué d'enregistrements sonores, de transcriptions enrichies d'annotations sous le logiciel ELAN et d'informations sur les participants.

Méthodologie

La phase de collecte du corpus TuLHis1 a été réalisée avec le dispositif du *Laboratoire Mobile des langues* (https://ecouter-parler.fr/). Ce dispositif a permis d'accéder à deux types de terrains : au pied de la médiathèque d'Orléans dans le cadre du Festival *Ozélir!* organisé par la médiathèque départementale du Loiret en mai 2023 et dans une guinguette de bord de Loire à Orléans (*La Paillotte*). Une interface de gestion des enregistrements et de métadonnées est intégrée au dispositif.

Un premier traitement des enregistrements a consisté en un découpage des fichiers sons afin d'isoler la partie purement lecture des parties « pré-lecture » (les échanges concernant le choix du livre), et « post-lecture » (les échanges de commentaires sur le livre).

La réalisation des transcriptions et annotations s'est faite en plusieurs phases : une première version de transcription avec l'outil de transcription automatique *Whisper*, disponible dans le service *Sharedocs* de l'IR* Huma-Num (modèles *Large* et *With speaker*), puis une phase de relecture avant les annotations. Deux types d'annotations ont été réalisées : annotation des liaisons et annotation des parties de l'album.

Concernant les annotations des liaisons nous avons repris une procédure déjà mise en place pour d'autres projets (Dugua et al. 2022 pour les détails). Après un alignement mot/son par le logiciel *Jtrans* (Cerisara et al., 2009), compilée à la segmentation initiale en tours de parole, nous passons un étiquetage automatique en POS avec *Treetagger* via l'outil *TEIcorpo*. Un repérage automatique des contextes de liaisons est ensuite réalisé sur la base d'une définition graphique de la liaison. On recherche des suites de deux mots dans lesquelles le premier se termine par l'une de ces lettres : « n, s, z, x, t, d, r, p » et le suivant commence par une voyelle

ou un « h ». A cette définition très générale, nous avons exclu certains mots que l'on sait ne pas entrer en contexte de liaison, par exemple : « non » comme Mot1, « oui » et « ouais » comme Mots2, etc. Une préannotation automatique attribue la consonne de liaison selon la lettre finale du premier mot : /n/ pour « n », /z/ pour « s », « x », /t/ pour « t », « d », etc. Une fois les contextes de liaisons pré-annotés, nous réécoutons pour valider ou modifier la pré-annotation avec le logiciel *ELAN*.

Une deuxième couche d'annotation rend compte des différentes parties qu'on peut trouver dans album (voir Dugua, 2023 pour la grille d'annotation), à savoir par exemple les passages dialogués, les descriptions, les discours rapportés, etc. Cette couche d'annotation permet également d'identifier les passages qui ne sont pas de la lecture mais des échanges entre l'enfant et l'adulte.

Résultats

Le corpus étant en cours de traitement, les données ne sont pas encore disponibles pour rendre compte des résultats. Nous pouvons toutefois nous appuyer sur trois tendances qu'une étude préliminaire (Dugua et al. 2022) avait mises en évidence.

La première concernait l'influence des parties d'albums. En ne prenant en compte que les temps de lecture, nous avions pu mettre en évidence une différence dans l'usage des liaisons entre les temps de dialogue (au sein de l'histoire) et les temps de narration, et plus précisément, entre les dialogues pris en charge par des enfants et ceux produits par des adultes. Comme si le lecteur, projetait dans sa lecture le fait que les enfants pourraient faire moins de liaison que des adultes.

La deuxième tendance s'appuyait sur le critère de l'âge de l'enfant qui semblait être un facteur déterminant de l'usage des liaisons variables en lecture par l'adulte. Ces dernières étaient plus fréquemment réalisées lorsque la lecture s'adressait à des enfants de 5-6 ans qu'à des enfants de 2-3 ans. A l'image de ce qu'on sait de l'évolution du CDS en parole spontanée, il semble qu'on retrouve une évolution en lecture. Dans TuLHis1, avec des données quantitativement plus importantes, nous pourrons constituer plusieurs groupes d'enfants selon leurs âges pour vérifier cette tendance et observer de manière plus fine cette évolution, savoir par exemple si elle est progressive ou s'il y a des paliers.

Une dernière tendance concernait la nécessité de croiser les différents facteurs sociolinguistiques pour rendre compte des usages de la liaison avec également des facteurs plus fins comme le lien entre le lecteur et l'enfant ou le type d'album. Ici, notre panel de lecteurs et lectrices sera plus important et devrait permettre d'affiner ces éléments.

Ce papier se focalisera sur deux aspects primordiaux de la linguistique de corpus : la mise en place d'un projet de constitution d'un corpus oral et l'extraction de premiers résultats avec une approche quantitative (des taux de liaisons réalisées) croisée d'observations plus qualitatives sur des contextes de liaisons particuliers ou en décrivant des caractéristiques des locuteurs.

Références bibliographiques

Cameron-Faulkner, T., & Hickey, T. (2011). Form and function in Irish child directed speech. *Cognitive Linguistics*, 22(3), 569-594.

Cerisara, C., Mella, O., & Fohr, D. (2009, September). JTrans, an open-source software for semi-automatic text-to-speech alignment. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association-Interspeech 2009*.

Coutanson, G. (2023). *Pataquès et liaison : étude de deux phénomènes de sandhi externe dans des corpus de français chanté*, Thèse de doctorat, Université Paris Nanterre.

Dugua, C., Badin, Fl., Fallon, B. & 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.

Dugua, C. (2023). Usage des liaisons variables dans deux corpus de lecture. *Langue française*, 219 : 17-32.

Fortier, N. (2023). *Histoires de Lou*, Editions du Rouergue.

Foulkes, P., Docherty, G., & Watt, D. (2005). Phonological variation in child-directed speech. *Language*, 81, 177-206.

Frier, C. (Éd.). (2006). *Passeurs de lecture : Lire ensemble à la maison et à l'école*. Paris : Retz.

Gadet, F. (1989). *Le français ordinaire*. Armand Colin.

Grossmann, Fr. (2006), Logiques sociales et clivages culturels dans les lectures partagées. In Frier, C. (ed.), *Passeurs de lectures – lire ensemble à la maison et à l'école*. Paris : Retz, 19-43

Langlard, H. (1928). *La liaison dans le français*. Librairie ancienne Edouard Champion.

Snow, C. E. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, 4(1), 1-22.

Snow, C. E. (1999). Beginning from baby talk : Twenty years of research on input in interaction. In C. Gallaway & B. J. Richards (Éds.), *Input and interaction in language acquisition*. Cambridge University Press, 1-12.

Vers une étude de la perception des changements urbains

Aurore Lessieux [1], Iris Eshkol-Taravella [1] et Olivier Ratouis [2]
[1] Laboratoire MoDyCo, Université Paris Nanterre
[2] Laboratoire LAVUE, Université Paris Nanterre
alessieux@parisnanterre.fr, ieshkolt@parisnanterre.fr, oratouis@club-internet.fr

Résumé

Ce travail s'intéresse à la manière dont les différents acteurs urbains perçoivent un projet d'aménagement, à partir d'un corpus oral et multimodal composé de journaux télévisés, de reportages, d'interviews et d'émissions de radio issus des bases de données de l'INA. Cette communication s'articule autour des enjeux et des étapes de constitution d'un corpus permettant l'étude de la perception des changements urbains. Elle s'ouvre sur la définition de l'objet d'étude dans une perspective pluridisciplinaire, puis se poursuit par la présentation de la méthodologie adoptée pour constituer un corpus brut et annoté, conçu en cohérence avec les objectifs du projet.

La perception des changements urbains

Dans son ouvrage *La réception sociale de l'urbanisme* (2007), Nora Semmoud soutient que les projets urbains ne peuvent être pleinement compris sans considérer la manière dont ils sont perçus, appropriés et parfois détournés par les usagers. Ses travaux sur la réception des projets urbains constituent un apport fondamental pour penser la manière dont les transformations de l'espace sont comprises, ressenties et jugées par les acteurs concernés. À partir d'enquêtes empiriques menées notamment en Algérie et en France, Semmoud met en évidence que tout projet d'aménagement fait l'objet d'appropriations différenciées, de résistances, de reconfigurations symboliques. Son approche s'attache à la manière dont ces projets sont perçus et appropriés, révélant ainsi la pluralité des points de vue, des émotions et des représentations sociales liées à l'espace. Si la notion de réception met en lumière le rapport dialogique entre un projet urbain et son public, elle repose souvent sur une vision binaire émetteur/récepteur, suggérant une posture passive du second, mais elle peine à rendre compte de la complexité subjective de cette interaction. C'est la raison pour laquelle nous lui préférons le terme de perception, qui insiste sur le caractère actif, interprétatif et incarné de l'expérience urbaine. Percevoir ne se limite pas à recevoir un message : c'est une construction de sens, fondée sur des évaluations cognitives, émotionnelles et sociales. Appliqué aux projets d'aménagement, cela signifie que les réactions des habitants, usagers ou experts s'inscrivent dans un ensemble complexe de représentations spatiales, d'attentes, de souvenirs, d'attachements.

Constitution du corpus

La présente contribution se concentre sur le projet EuropaCity, en raison de la forte controverse publique que ce projet a suscitée, révélant des divergences marquées de perception entre promoteurs, élus, habitants et opposants. Projet de complexe urbain mêlant loisirs, commerces et espaces culturels sur le site du Triangle de Gonesse, EuropaCity a cristallisé des tensions

médiatisées jusqu'à son abandon en 2019, offrant ainsi un terrain d'étude particulièrement riche pour observer les dynamiques discursives autour d'un projet urbain controversé.

Pour constituer le corpus d'étude, nous avons mobilisé les ressources de l'INA, qui offre un accès à des contenus médiatiques archivés et finement indexés. Cette base permet de travailler sur des discours publics authentiques, produits dans des contextes réels et impliquant une grande diversité de locuteurs. La sélection du corpus s'est appuyée sur plusieurs critères. Il devait, d'une part, porter sur un ou plusieurs projets urbains clairement identifiés et ayant suscité une couverture médiatique substantielle. D'autre part, il était essentiel que les documents soient disponibles dans les bases de données de l'INA, condition préalable à leur exploitation. Le corpus devait également présenter un caractère polémique ou controversé, afin de faire émerger des prises de position contrastées dans l'espace public. Enfin, une attention particulière a été portée à la diversité des profils de locuteurs, incluant élus, urbanistes, usagers, experts, militants, et autres acteurs aux degrés d'implication variés.

Le corpus rassemble 281 documents audiovisuels (environ 17 heures), diffusés entre 2012 et 2022. Pour y parvenir plusieurs étapes furent nécessaires en commençant dans un premier temps par la sélection du corpus qui a consisté à recueillir les documents audiovisuels dans les fonds de l'INA. Cette étape a débuté par l'organisation de réunions pluridisciplinaires réunissant urbanistes, linguistes et documentalistes, afin d'évaluer la faisabilité d'un travail sur le projet EuropaCity. Ces échanges ont également permis un partage de nos vocabulaires spécialisés respectifs, ainsi qu'une prise en main des fonctionnalités de l'outil de requête disponible à l'INAthèque pour explorer les fonds audiovisuels. Les requêtes permettent d'obtenir des listes de notices documentaires correspondant aux critères de recherche, et c'est sur les informations qu'elles contiennent que s'est fondé le choix d'inclure ou non chaque document dans le corpus. Chaque document audiovisuel est accompagné d'une notice documentaire rédigée par les documentalistes de l'INA, indiquant notamment les métadonnées relatives à la diffusion du média (titre, date, durée, etc.), des descripteurs, et des résumés. Différentes combinaisons de mots-clés ont été utilisées pour effectuer les requêtes : le nom du projet seul ou associé à un type de contenu (interview, micro-trottoir), l'association d'une dénomination géographique du lieu avec un vocabulaire lié à la polémique ou à l'aménagement urbain, etc. Les notices issues de ces requêtes ont ensuite été rassemblées, puis triées manuellement en fonction de leur pertinence pour le projet. Plusieurs critères ont été pris en compte : le sujet principal porte-t-il effectivement sur EuropaCity, ou s'agit-il simplement d'une mention secondaire (exemple, nom de fonction, localisation) ? Quelle place (en termes de durée) est accordée au projet ? La diversité des profils de locuteurs est-elle représentée ?

Dans un second temps, il a été nécessaire de prétraiter le corpus, les données brutes étant inexploitables en l'état ; leur conversion en format textuel constituait en effet une condition préalable à l'application des outils TAL. La transcription automatique des documents audiovisuels a été réalisée par l'INAlab, qui permet l'exploitation à grande échelle des collections de l'INA et propose divers outils de fouille et d'analyse automatisée, spécifiquement adaptés aux corpus multimédias, notamment plusieurs outils de reconnaissance de la parole. L'outil WhisperX a été retenu en raison de la qualité de sa segmentation en tours de parole et de la précision de son alignement entre la transcription et l'audio. Les transcriptions étant au format JSON, plusieurs scripts Python ont été développés afin de produire des formats de données interopérables avec différents outils d'exploitation de corpus oraux (CLAN, ELAN) et d'annotation (Glozz). Les transcriptions ont fait l'objet d'une révision manuelle visant à affiner la segmentation en tours de parole, à corriger les erreurs d'alignement et à rectifier les imperfections issues de la transcription automatique. Les erreurs de graphies des entités

nommées, les ambiguïtés phonologiques, les omissions liées aux bruits de fond, ainsi que les altérations des marques propres de l'oral ont également été corrigées.

Modèle de la Perception et Annotation manuelle

La modélisation de la perception proposée s'appuie sur un socle théorique composite, croisant des approches issues de la linguistique systémique fonctionnelle, de l'extraction d'opinions et de l'analyse des émotions (Martin & White 2005, Liu 2012, Etienne 2023). Elle rend compte de la diversité des manifestations linguistiques de la perception des projets d'aménagement dans les discours médiatiques et repose sur six dimensions permettant son annotation fine : 1. **Expérienceur** (*qui perçoit*) : identification du locuteur ou du groupe auquel est attribuée la perception en fonction de son implication dans le projet urbain ; 2. **Type de perception** : distinction entre *émotion*, *jugement* ou *intention*, selon la nature de perception exprimée sur la base d'indices linguistiques ; 3. **Polarité** : positive, négative ou neutre ; 4. **Intensité** : codage de l'intensité selon la présence d'indices linguistiques ; 5. **Ancrage temporel** : repérage de l'étape du projet à laquelle se rattache la perception ; 6. **Cible** : désignation de l'entité (le projet urbain) et de l'aspect (caractéristique du projet) visé par la perception.

Le corpus a fait l'objet d'une annotation manuelle selon le modèle de la perception proposé. Trois annotateurs ont été mobilisés, avec un accord inter-annotateur global de 0,65 (κ de Cohen ; Landis & Koch, 1977). Lorsqu'une perception relative au projet EuropaCity est exprimée dans un tour de parole, celui-ci fait l'objet d'une annotation. La détection de ces perceptions repose sur l'identification d'indices linguistiques, qu'ils soient verbaux (morphologiques, lexicaux, syntaxiques, stylistiques) ou non verbaux (prosodiques, paralinguistiques), qui sont systématiquement relevés et annotés.

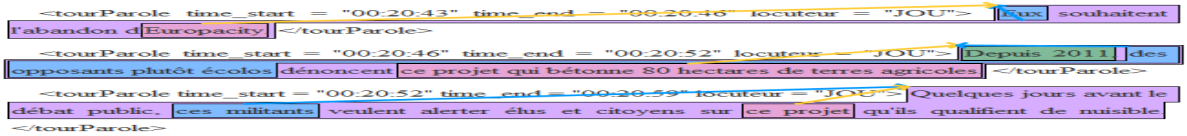


figure . 1 Exemple d'annotation des transcriptions dans Glozz

Dans cet exemple, le tour de parole comporte une perception de type **jugement**, à **polarité négative**, identifiable grâce à plusieurs indices verbaux tels que *opposants*, *dénoncent* et *bétonne 80 hectares de terres agricoles*. L'**expérienceur** est marqué par un segment *des opposants plutôt écolos*. L'**entité cible** de la perception est le projet EuropaCity et l'**aspect ciblé** est *la bétonisation de 80 hectares de terres agricoles*, une conséquence du projet urbain. Enfin, l'expression *depuis 2011* constitue l'**ancrage temporel** de la perception, correspondant à une phase antérieure au développement du projet urbain.

L'outil d'annotation retenu pour cette étude est le logiciel **Glozz**. Bien qu'il ait initialement été conçu pour l'annotation de corpus écrits, il présente plusieurs fonctionnalités particulièrement adaptées à notre modèle d'annotation. Glozz permet en effet d'associer différents attributs à chaque étiquette, de structurer une annotation hiérarchique, de superposer les annotations, et de définir des relations entre celles-ci. Ces caractéristiques ont motivé un choix en faveur de Glozz, au détriment d'autres outils comme INCEpTION. Ce dernier, bien que récemment amélioré pour répondre à des besoins similaires, présente encore certaines limitations notables, telles que la suppression des balises XML, utilisées ici pour structurer les tours de parole et maintenir l'alignement avec l'audio, ou encore la gestion peu ergonomique des relations dans des corpus de grande taille. Il est à noter toutefois que des travaux en cours visent à rendre Glozz et INCEpTION interopérables.

Les résultats issus de l'annotation manuelle du corpus permettront de mettre en lumière les régularités discursives, les formes d'expression de la perception, ainsi que les dynamiques énonciatives propres aux différents acteurs urbains.

Références bibliographiques

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. In Proceedings of Interspeech 2023.

Etienne, A. (2023). Analyse automatique des émotions dans les textes : Contributions théoriques et applicatives dans le cadre de l'étude de la complexité des textes pour enfants [Thèse de doctorat, Université Paris Nanterre – Paris X].

Flamein, H., & Eshkol-Taravella, I. (2020). De la parole à la carte : repérage, analyse et visualisation automatique de la perception d'une ville. SHS Web of Conferences.

Martin, J. R., & White, P. R. R. (2005). The Language of Evaluation: Appraisal in English. New York : Palgrave Macmillan.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.

Semmoud, N. (2007). La réception sociale de l'urbanisme. Paris : L'Harmattan.

Semmoud, N. (2015). Les marges urbaines : un analyseur privilégié de l'urbanisme d'Alger ? Les Cahiers d'EMAM, (27).

Widlöcher, A., & Mathet, Y. (2012). The Glozz platform: A corpus annotation and mining tool. In Proceedings of the 2012 ACM Symposium on Document Engineering (pp. 171–180). ACM.

Vers une typologie des erreurs de reprise anaphorique dans des rédactions étudiantes

Vanessa Gaudray Bouju 1, 2, Iris Eshkol-Taravella 1, 2 et Sabine Lehmann 1, 2
1 MoDyCo, Université Paris Nanterre, 2 écri+
v.gaudraybouju@parisnanterre.fr, ieshkolt@parisnanterre.fr, slemhajeb-lehmann@parisnanterre.fr

Introduction

Les anaphores sont des expressions référentielles décrites classiquement comme reprenant un segment antérieur du discours. Leur interprétation s'appuie sur leur antécédent mais aussi sur le cotexte, impliquant des mécanismes sémantiques et cognitifs complexes (Johnsen, 2019). Du point de vue discursif, elles servent à reprendre l'information textuelle afin d'assurer la continuité thématique du propos. Elles sont donc essentielles à la cohésion textuelle et à la dynamique informationnelle du texte (Charolles, 1995). De fait, il est important de bien les manipuler afin de produire un texte clair et cohérent. Or, nous observons dans les rédactions étudiantes que leur usage est source de difficultés. En outre, la notion de reprise anaphorique est vaste et fait appel à des règles de différents niveaux linguistiques (grammaire, sémantique, pragmatique), que même les locuteurs avancés peuvent avoir des difficultés à maîtriser (Reichler-Béguelin et al., 1990). Ces constats justifient l'intérêt porté à ce phénomène.

Si différents auteurs, sur les travaux desquels nous reviendrons, ont déjà relevé et analysé certains emplois déviants des reprises anaphoriques, aucune étude ne condense l'ensemble de ces travaux pour proposer une typologie des problèmes rencontrés. La recherche menée ici, financée par le projet écri+[1], s'inscrit dans le domaine de la linguistique de corpus et porte sur les littéracies universitaires (Delcambre & Lahanier-Reuter, 2010), le corpus d'étude constitué étant composé de rédactions étudiantes. Cette recherche vise à développer une ressource annotée en problèmes de reprises afin d'entraîner un outil de détection automatique. Pour ce faire, une démarche empirique a été entamée afin de proposer une typologie rigoureuse et homogène des emplois déviants de reprises anaphoriques, fondée sur les critères formels, qui sera utilisée pour l'annotation manuelle préalable à tout traitement automatique. Après avoir fait un bref état de la littérature sur ces questions et présenté le corpus d'étude ainsi que la méthodologie suivie, nous proposons une typologie des problèmes observés.

État de l'art

Dans ce travail, nous nous intéressons aux emplois déviants des reprises anaphoriques. Nous serons amenés à parler d'erreur dans un sens large, en nous inspirant des réflexions faites par Anctil dans le cadre de sa thèse portant sur les erreurs lexicales (2011), qui rejette l'opposition binaire entre erreur et maladresse pour parler d'erreurs de différents niveaux de « gravité ».

1 Le projet ANR écri+ se dote de ressources pédagogiques variées afin d'aider les étudiants à développer leurs compétences rédactionnelles.

Plusieurs auteurs se sont intéressés aux erreurs ou autres emplois déviants dans l'usage des reprises anaphoriques dans les rédactions étudiantes. Reichler-Béguelin (1988) analyse ces dysfonctionnements à travers le principe de coopération entre le locuteur et le récepteur, le locuteur étant censé s'appuyer sur leur savoir partagé pour produire des anaphores interprétables. Toutefois, un élément saillant dans l'esprit du scripteur ne l'est pas toujours dans le discours, ce qui peut générer des difficultés d'interprétation. En outre, le problème ne vient pas forcément de l'impossibilité de trouver le référent mais du coût de décodage de l'énoncé jugé excessif, ce qui le rend sanctionnable. Lang et Meyer (2015) ont pour leur part réalisé une étude croisée entre les erreurs d'accord et de reprise anaphorique dans les rédactions étudiantes. Ils ont montré que les deux phénomènes peuvent être appréhendés comme des chaînes dont certains maillons sont moins solides et donc susceptibles d'entraîner des erreurs. Ils relèvent notamment l'importance de deux paramètres : l'iconicité (éloignement ou proximité du référent dans le texte) et la conceptualisation (représentation cognitive du référent).

Les causes pouvant expliquer les erreurs rencontrées sont multiples. Schnedecker (1995) a notamment montré que l'enseignement des reprises à l'école primaire et les commentaires des professeurs se focalisent essentiellement sur le fait d'éviter les répétitions. Les stratégies mises en place par les élèves pour éviter à tout prix les répétitions peuvent être source de maladresses rompant la continuité référentielle, en plus de ne pas être nécessaires dans certains contextes.

Des ébauches d'énumération ou de classification des problèmes rencontrées dans l'usage des reprises anaphoriques ont été proposées par ces auteurs. Reichler-Béguelin (1988) distingue les erreurs d'ordre linguistique (choix des expressions anaphoriques, accord avec l'antécédent) et les erreurs d'ordre pragmatique (disponibilité du réfèrent dans la mémoire discursive). Schnedecker (1995) a pour sa part étudié les chaînes de reprises dans des textes argumentatifs produits par des élèves. Ce genre manipule des objets discursifs plus abstraits et changeants que le narratif, pouvant prendre des formes grammaticales variées. L'auteur relève quatre principaux problèmes qui en découlent : une confusion possible entre des référents abstraits ou particularisants ; un mauvais usage du démonstratif pour fusionner des référents disjoints ; une mauvaise gestion de l'alternance entre reprises objectives ou avec éclairage subjectif ; enfin, une difficulté à conceptualiser des informations se traduisant par des emplois flous de « cela » ou de termes vagues. L'étude présentée ici s'inspire de ces différentes considérations pour proposer une typologie des erreurs de reprise se voulant la plus exhaustive possible.

Corpus et méthodologie

Le corpus de rédactions étudiantes constitué pour cette étude est composé de 360 textes produits dans le cadre universitaire par des étudiants en L2 d'éco-gestion. Ces rédactions assez brèves (environ une page chacune) sont issues de consignes rédactionnelles variées (présentation d'une notion, description d'un document, argumentation…) et forment un total de 100 000 mots. Ces textes ont été étudiés de façon qualitative : les énoncés contenant des reprises problématiques ont été extraits afin d'être analysés. Cette première passe manuelle a permis de faire émerger des similitudes entre différents types de problèmes, qui ont été regroupés.

Ces premières tentatives de classification ont toutefois mis en lumière plusieurs usages se situant à la frontière de l'acceptable. Pour questionner la réception de ces différents emplois, un formulaire a été adressé à 80 participants au profil varié (enseignants, étudiants, extérieurs au monde académique…). Les 21 énoncés soumis ont été choisis par la doctorante experte de la tâche, à partir des observations du corpus, pour rendre compte des différents types d'emplois de reprises pour lesquelles le jugement d'acceptabilité est problématique.

Les réponses des participants ont montré que les problèmes de reprise se cumulent souvent avec d'autres. Plus l'énoncé est globalement mal formé, plus les jugements à l'encontre de la reprise sont sévères. Notons que la compréhensibilité de l'énoncé n'est pas toujours le critère le plus important ; les répétitions superflues (comme en [3] ci-dessous) sont par exemple assez largement sanctionnées. Concernant les reprises ambiguës, le fait que le contexte permette de les désambiguïser n'est pas suffisant, les « fausses pistes » étant mal reçues. Au contraire, les reprises résomptives ou conceptuelles floues (e.g. « cela », « en ») sont peu sanctionnées quand le sens global de l'énoncé est compréhensible. Ces différents constats ont permis de rapprocher à nouveau certains types de problèmes en se basant sur le critère de la réception des reprises.

Vers une typologie des erreurs

À partir de l'ensemble de ces observations, une première ébauche de typologie est amorcée. Nous présentons ici les trois principales catégories d'erreurs ayant retenu notre attention.

Les erreurs grammaticales regroupent les cas où les règles de grammaire ne sont pas respectées dans le procédé de reprise, ce qui rend l'énoncé incorrect grammaticalement parlant. Les erreurs d'accord (en genre ou en nombre) entre la reprise et son antécédent, comme en [1], constituent le cas le plus représentatif de cette catégorie.

> [1] Chaque pays détient des spécialisations leur permettant de créer de la richesse.

Une seconde catégorie regroupe les problèmes liés à l'antécédent et à son identification. Elle peut être rapprochée de ce que Reichler-Béguelin (1988) appelle erreurs d'ordre pragmatique, liées à la disponibilité du référent dans la mémoire discursive. Nous distinguons les difficultés liées au manque d'accessibilité de l'antécédent (ambiguïté due à une compétition entre plusieurs référents potentiels, comme en [2] ; distance trop importante avec l'antécédent ; etc. ; voir les facteurs d'accessibilité d'Ariel (1988)) des cas, plus problématiques, où aucun antécédent n'est identifiable ou récupérable à l'aide d'inférences simples.

> [2] La prohibition de cette substance n'a sans doute pas été la meilleure chose à faire. On a pu voir qu'elle doit tout de même être limitée comme toutes les drogues présentes dans notre société.

Enfin, une troisième catégorie regroupe les cas où le problème vient directement de la reprise du point de vue lexical. Elle concerne ce qu'Apothéloz (1995) nomme l'acte de dénomination, distingué de l'acte de référence, tous deux inclus dans le processus de reprise. Les problèmes concernés peuvent consister en des cas de *répétitions* ou *redondance*, i.e. en l'emploi de marqueurs de faible accessibilité (sémantiquement riches) quand l'antécédent est déjà saillant, comme en [3], où on attendrait plutôt un marqueur de forte accessibilité tel qu'un pronom (voir la typologie des marqueurs d'accessibilité d'Ariel (1988) et la hiérarchie de la donation de Gundel *et al.* (1993)). Le terme choisi peut également ne pas convenir s'il est porteur de flou ou d'imprécision lexicale.

> [3] J'aime énormément les livres qui nous apportent une leçon, car chaque histoire a une leçon.

La définition de ces catégories sert de support à la rédaction d'un guide d'annotation, qui permet de guider la création d'un corpus annoté en erreurs de reprise. Ce corpus donnera la possibilité de passer à une autre échelle et de réaliser des analyses quantitatives autour de ce phénomène. À terme, l'ambition de ce projet est d'utiliser le corpus annoté avec la typologie présentée ici pour entraîner un outil de détection automatique des problèmes dans l'usage des reprises. Cet

outil, à visée pédagogique, entend être intégré au projet écri+ pour aider les étudiants à corriger leurs textes par eux-mêmes.

Références bibliographiques

Anctil, D. (2011). L'erreur lexicale au secondaire : analyse d'erreurs lexicales d'élèves de 3e secondaire et description du rapport à l'erreur lexicale d'enseignants de français [thèse de doctorat, Université de Montréal].

Apothéloz, D. (1995). Rôle et fonctionnement de l'anaphore dans la dynamique textuelle (Vol. 29). Librairie Droz.

Ariel, M. (1988). Referring and accessibility. Journal of Linguistics, 24(1), 65–87.

Delcambre, I., & Lahanier-Reuter, D. (2010). Les littéracies universitaires: influence des disciplines et du niveau d'étude dans les pratiques de l'écrit. Forumlecture. ch, 3, 1-17.

Charolles, M. (1995). Cohésion, cohérence et pertinence du discours. Travaux de Linguistique : Revue Internationale de Linguistique Française, 1995, 29, pp.125-151.

Gundel, J.K., Hedberg, N. & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. Language 62(2), 274–307.

Johnsen L.A. (2019.) La sous-détermination référentielle et les désignateurs vagues en français contemporain [thèse de doctorat, Université de Neuchâtel].

Lang, É., & Meyer, J. P. (2015). Grammaire avancée et littéracies universitaires. Vos papiers sont-ils en règle. Les études françaises aujourd'hui. Pourquoi étudier la grammaire, 223-242.

Reichler-Béguelin, M.-J. (1988). Anaphore, cataphore et mémoire discursive. In: Pratiques : linguistique, littérature, didactique, n°57, 1988. L'organisation des textes. pp. 15-43.

Reichler-Béguelin, M. J., Denervaud, M., & Jespersen, J. (1990). Écrire en Français : cohésion textuelle et apprentissage de l'expression écrite.

Schnedecker, C. (1995). Besoins didactiques en matière de cohésion textuelle : les problèmes de continuité référentielle. In: Pratiques : linguistique, littérature, didactique, n°85, 1995. pp. 3-25.