La linguistique dans une ère nouvelle : discours, méthodes et technologies dans le paysage contemporain

Vanessa Gaudray Bouju, Ho Won Kim, Charlotte Laffargue, Xibin Wang, Santiago Herrera

▶ **To cite this version:**

Vanessa Gaudray Bouju, Ho Won Kim, Charlotte Laffargue, Xibin Wang, Santiago Herrera. La linguistique dans une ère nouvelle : discours, méthodes et technologies dans le paysage contemporain. COLDOC, 2024. hal-04819687

HAL Id: hal-04819687
https://hal.science/hal-04819687v1

Submitted on 4 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

ColDoc 2024

14 – 15 Octobre 2024

Université Paris Nanterre

Bâtiment de la formation continue

Acte du Colloque

de la 15e édition du Coldoc

La linguistique dans une ère nouvelle : discours, méthodes et technologies dans le paysage contemporain

Invités

Patrick Charaudeau (Université Sorbonne Paris Nord / Cerlis Paris Cité)
Albin Wagener (Université Catholique de Lille / ESSLIL / Ethics EA 7446)
Emmanuel Ferragne (Université Paris Cité / IUF / CLILLAC-ARP / Labex EFL)
Marie-Catherine de Marneffe (UCLouvain / CENTAL)

modyco Modèles, Dynamiques, Corpus UMR 7114

Université Paris Nanterre UFR PHILLIA

Université Paris Nanterre

ATALA

CORLI

ColDoc 2024

Programme : Lundi 14 Octobre

9h-9h15	Accueil des participants
9h15-10h15	**[Mot d'ouverture] Patrick Charaudeau :** « Une ère nouvelle en sciences du langage ? Ce qu'est une discipline »

Session 1 : Analyse de discours de communication : presse et militantisme

Présidée par Sabine Lehmann et Caroline Facq-Mellet

10h15-10h45	**Yoohee Yoon** : « Analyse sémantique du nom violence dans le discours médiatique : l'émergence des phénomènes sociaux de violences policières et des violences faites aux femmes »
10h45-11h15	**Xuemin Li** : « Comparaison des opérations de reformulation dans des revues de vulgarisation scientifique en français et en chinois »
11h15-11h30	Pause café
11h30-12h30	**[Conférencier invité] Emmanuel Ferragne :** « Phonétique et phonologie « de corpus » : quelques notes épistémologiques »
12h30-14h	Déjeuner
14h-15h	Session Posters
15h-15h30	**Louis Escouflaire** : « La voix derrière le masque : méthode hybride d'analyse de la subjectivité dans le discours de presse »
15h30-16h30	**[Conférencier invité] Albin Wagener** : « Explorer les controverses postdigitales grâce à l'analyse de discours : des féminazis aux écoterroristes »
16h30-16h45	Pause café

Session 2 : Discours dans l'espace numérique

Présidée par Anne Lacheret-Dujour, Hugo Dumoulin et Santiago Herrera

16h45-17h15	**Liudmila Oshchepkova** : « Expression de l'évaluation dans le commentaire sur Internet »
17h15-17h45	**Mohamed Araci** : « L'usage des mèmes par les utilisateurs algériens sur Facebook : provoquer le rire et refléter la vie sociale. »
20h-23h	Dîner au restaurant Le Polidor à Paris

ColDoc 2024

Programme : Mardi 15 Octobre

9h-9h45	Petit Déjeuner
9h45-10h15	**Trang Pham Tran Hanh, Rémi Cardon, Rodrigo Wilkens, Thomas François** : « Analyse du genre de discours dans des sites web des agences de voyages francophones »
10h15-10h45	**Marco Antonio Almeida Ruiz** : « Technologie et société dans le monde postpandémique : les réseaux numériques dans la construction de mémoires et d'évènements discursifs dans l'envers de l'histoire »
10h45-11h	Pause café
11h-12h	**[Conférencière invitée] Marie-Catherine de Marneffe :** « Au-delà de l'étiquette unique: Vers une annotation plus riche en TAL »

Session 3 : Corpus d'aujourd'hui

Présidée par Iris Eshkol-Taravella et Sylvain Kahane

12h-12h30	**Anaïs Dagniaux** : « Textométrie et corpus oral : apports et limites pour une analyse des effets discursifs du vieillissement »
12h30-14h	Déjeuner
14h-15h	Session Poster
15h-15h30	**Rayan Ziane, Fatma Ben Barka Messaoudi** : « Bootstrapper son corpus oral ou comment développer un corpus d'arabe parlé dans le petit Maghreb »
15h30-16h	**Althea Löfgren, Natalia Levshina** : « Testing trade-offs between gender and number indexing and other cues to A and P arguments: A corpus-based perspective »
16h-16h15	Pause café
16h15-16h45	**Aurore Lessieux** : « Typologie de la perception des dynamiques urbaines dans des données multimodales : cas du projet d'Europacity »
16h45-17h15	**Adam Faci, Antoine Silvestre De Sacy** : « RAG pour l'exploration de corpus en GLAM »

Conférenciers invités

Charaudeau Patrick (Université Sorbonne Paris Nord / Cerlis Paris Cité)

De Marneffe Marie-Catherine (UCLouvain / CENTAL)

Ferragne Emmanuel (Université Paris Cité / IUF / CLILLAC-ARP / Labex EFL)

Wagener Albin (Université Cathollique de Lille / ESSLIL / Ethics EA 7446)

Comité scientifique

Balvet Antonio (Université de Lille)

Bastitelli Delphine (Université Paris Nanterre)

Beillet Marie (Université Paris Nanterre)

Bogliotti Caroline (Université Paris Nanterre)

Cislaru Georgeta (Université Paris Nanterre)

Charaudeau Patrick (Honorifique. Université Sorbonne Paris nord / Cerlis Paris Cité)

Claudel Chantal (Université Paris Nanterre)

Désagulier Guillaume (Université Bordeaux-Montaigne)

Doury Marianne (Université Paris Descartes)

Dumoulin Hugo (Université Paris Nanterre)

Eshkol-Taravella Iris (Université Paris Nanterre)

Evrard Marc (Université Paris-Saclay)

Guibon Gaël (Université de Lorraine)

Guillaume Bruno (INRIA)

Heidlmayr Karin (Université Paris Nanterre)

Kahane Sylvain (Université Paris Nanterre)

Kraif Olivier (Université Grenoble Alpes)

Lampitelli Nicolas (Université Paris Nanterre)

Lehmann Sabine (Université Paris Nanterre)

Loock Rudy (Université de Lille)

Mahé Gwendoline (Université de Lille)

Novakova Iva (Université Grenoble Alpes)

Paveau Marie-Anne (Université Sorbonne Paris Nord)

Piccoli Vanessa (Université Paris Nanterre)

Prévost Philippe (Université de Tours)

Rakotonoelina Florimond (Université Sorbonne Nouvelle)

Reboul-Toure Sandrine (Université Sorbonne Nouvelle)

Santiago Fabian (Université Paris 8 Vincennes-Saint Denis)

Sitri Frédérique (Université Paris-Est Créteil)

Talbot Aurélien (Université Grenoble Alpes)

Tellier Marion (Université Aix Marseille)

Villoing Florence (Université Paris Nanterre)

Wagener Albin (Université Catholique de Lille)

Comité d'organisation

Gaudray Bouju Vanessa (Université Paris Nanterre)

Herrera Santiago (Université Paris Nanterre)

Kim Ho Won (Université Paris Nanterre)

Laffargue Charlotte (Université Paris Nanterre)

Wang Xibin (Université Paris Nanterre)

Financement

ATALA

Université Paris Nanterre UFR PHILLIA

modyco Modèles, Dynamiques, Corpus UMR 7114

CORLI

Sommaire

APPEL A SOUMISSION

COLDOC 2024 - Colloque de doctorants et jeunes chercheurs

MoDyCo UMR 7114 CNRS
Université Paris Nanterre
14 et 15 octobre 2024

"La linguistique dans une ère nouvelle : discours, méthodes et technologies dans le paysage contemporain."

ColDoc est un colloque international bisannuel organisé par le laboratoire MoDyCo (UMR 7114 - CNRS/Université Paris Nanterre) destiné aux jeunes chercheurs et doctorants en linguistique.

Pour sa 15ème édition, le **ColDoc** propose d'aborder une vision d'ensemble de la linguistique dans un contexte de changements technologiques. Il existe en effet des relations étroites entre la linguistique, les données langagières, au sens large, et les nouvelles technologies, au point que celles-ci paraissent aujourd'hui indissociables. Cette mise en lien est d'ordre multiple et fait partie de la recherche en linguistique.

L'ouverture de nouveaux champs d'étude en linguistique a conduit à la mise au point d'outils d'élaboration, de recueil et de traitement de différentes données.

L'étude de langues peu dotées ou décrites, par exemple, a besoin de l'adaptation d'outils de traitement automatique du langage (TAL), de traitement du signal, ou de recueil de données audiophonologiques (EGG/EMA/ultrasons), que ce soit pour l'analyse automatique, ou pour assister les linguistes de terrain (Bird, 2022 ; Ponti 2019). Le travail avec des corpus volumineux ou composés de divers genres ou langues nécessite de nouvelles méthodes statistiques d'analyse pour être représentatif (Levshina, 2019).

De la même manière, la recherche en psycholinguistique sur des sujets comme le plurilinguisme, l'acquisition et le vieillissement du langage (Bogliotti et al., 2017), ou encore le langage pathologique (Zhao et al., 2022), met en jeu des données comportementales (outils de type e-Prime, mouvements oculaires) et électrophysiologiques (EEG/IRMf) pour étudier les différents processus cognitifs qui interviennent durant le traitement linguistique (Peyre et Ramus, 2023)

La technologie ne se limite toutefois pas à un rôle d'outil analytique ; elle constitue également un vecteur de production langagière.

Les nouveaux discours multimodaux sont générés, prennent forme et circulent dans les espaces numériques (réseaux sociaux). Dans le même temps, de nouveaux genres discursifs s'imposent par leur centralité dans le débat public et quotidien, mettant en lumière des thèmes tels que les violences faites aux femmes (Association Faire Face, 2018 ; Lapalus, 2015) ou encore l'urgence climatique (Parrenin et Vargas, 2020). La question du statut de ces discours et de leur genre se pose alors (Rakotonoelina et Reboul-Touré, 2020).
Enfin, avec l'ouverture au public des grands modèles de langue, notamment via la génération de texte (chatGPT, Gemini, etc.), mais également dans d'autres applications telles que la

traduction (DeepL, Google translate, etc.), les outils technologiques deviennent sources de données langagières et de corpus artificiels, constituant à leur tour un objet d'étude (pour la traduction, par exemple, voir Loock, Lechauguette et Holt, 2022).

Nous invitons doctorants et jeunes chercheurs à proposer des communications qui traitent ces différentes questions. Les travaux, quel que soit leur état d'avancement, pourront aborder les nouveaux discours et/ou mettre en valeur de nouveaux outils, méthodologies et cadres théoriques innovants.
Cette diversité de perspectives permettra à des spécialistes de domaines ou outils différents, tels que la psycholinguistique, l'analyse du discours, la modélisation, etc., ayant en commun un regard contemporain sur la recherche en linguistique, de se rencontrer et d'échanger.

Les doctorants et les chercheurs intéressés par le sujet doivent soumettre une proposition de communication au format Word de 2 pages maximum hors bibliographie (police Times New Roman, taille de police 12) via le formulaire sur le site du Coldoc : https://coldoc2024.sciencesconf.org/

- *Les présentations peuvent se faire sous forme de communications orales, de posters, ou de démonstrations.*
- *Les langues de soumission sont le français et l'anglais.*
- *Les propositions de communication doivent être anonymes. Veuillez en outre préciser s'il s'agit d'une proposition pour une communication orale ou pour un poster.*
- ***La date limite de soumission est fixée au 15 mai 2024.***

Références

Bird, S. (2022). Local Languages, Third Spaces, and other High-Resource Scenarios. In S. Muresan, P. Nakov, & A. Villavicencio (Éds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* (p. 7817-7829). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.539

Bogliotti, C., Lacheret-Dujour A. & Isel, F. (2017) Atypies langagières de l'enfance à l'âge adulte. Apports de la psycholinguistique et des neurosciences cognitives. De Boeck Supérieur Editions

Association Faire Face. (2018). Le traitement médiatique des violences faites aux femmes : Entre instrumentalisation et invisibilisation. *GLAD!. Revue sur le langage, le genre, les sexualités*, *04*, Article 04. https://doi.org/10.4000/glad.1020

Lapalus, M. (2015). Feminicidio / femicidio : les enjeux théoriques et politiques d'un discours définitoire de la violence contre les femmes. *Enfances, Familles, Générations*, (22), 85–113. https://doi.org/10.7202/1031120ar

Levshina, N. (2019). Token-based typology and word order entropy : A study based on Universal Dependencies. *Linguistic Typology*, *23*(3), 533-572. https://doi.org/10.1515/lingty-2019-0025

Loock, R., Léchauguette, S., & Holt, B. (2022). Dealing with the « Elephant in the Classroom » : Developing Language Students' Machine Translation Literacy. *Australian Journal of Applied Linguistics*, *5*(3), 118-134.

Parrenin, F., & Vargas, É. (2020). Biodiversité et changement climatique : Entre discours du spécialiste et discours vulgarisé. *Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires*, *15*, Article 15. https://doi.org/10.4000/cediscor.2817

Paveau, M-A. (2017) *L'analyse du discours numérique. Dictionnaire des formes et des pratiques.* Hermann.

Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., & Korhonen, A. (2020). *Modeling Language Variation and Universals : A Survey on Typological Linguistics for Natural Language Processing* (arXiv:1807.00914). arXiv. https://doi.org/10.48550/arXiv.1807.00914

Rakotonoelina, F., & Reboul-Touré, S. (2020). Analyse du discours et biodiversité : Pluridisciplinarité, interdisciplinarité et transdisciplinarité. *Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires*, *15*, Article 15. https://doi.org/10.4000/cediscor.3181

Williams, C. M., Peyre, H., & Ramus, F. (2023). Brain volumes, thicknesses, and surface areas as mediators of genetic factors and childhood adversity on intelligence. *Cerebral Cortex (New York, N.Y.: 1991)*, *33*(10), 5885-5895. https://doi.org/10.1093/cercor/bhac468

Zhao, J., Song, Z., Zhao, Y., Thiebaut de Schotten, M., Altarelli, I., & Ramus, F. (2022). White matter connectivity in uncinate fasciculus accounts for visual attention span in developmental dyslexia. *Neuropsychologia*, *177*, 108414. https://doi.org/10.1016/j.neuropsychologia.2022.108414

Session 1

Analyse de discours de communication :

presse et militantisme

Analyse sémantique du nom violence dans le discours médiatique : l'émergence des phénomènes sociaux de violences policières et des violences faites aux femmes

YOON Yoohee

MoDyCo, Université Paris-Nanterre

yoonyohee@gmail.com

Mots-Clés : violence, problèmes publics, discours médiatique, sémantique discursive

L'objectif de cette communication est d'analyser le sens et les usages du mot violence dans le discours médiatique où des phénomènes sociaux comme les violences faites aux femmes et les violences policières apparaissent comme des problèmes publics.

Le discours médiatique est le lieu où l'émergence de nouveaux problèmes publics (Dewey 2010 [1927]) peut être observée. En effet, quand une situation problématique surgit dans la société, les faits s'y rapportant sont relatés par la presse. Cette dernière diffuse ensuite des débats sur le sujet qui font intervenir différents acteurs sociaux : les institutions, les personnalités politiques ou encore, les individus directement concernés par ces faits. Les journalistes rendent également compte des événements organisés afin de lutter contre le phénomène en question. Ils se font aussi l'écho des résultats d'études menées sur le sujet et des mesures prises par l'autorité publique pour résoudre le problème.

En ce qui concerne les phénomènes de violences faites aux femmes et de violences policières, le nom abstrait violence contribue à l'émergence du problème public. En effet, le mot polysémique violence est mobilisé dans le discours médiatique aux différentes étapes de l'établissement du problème public. Par exemple, lorsqu'il s'agit de relater les faits de violence qui ont concrètement eu lieu, le nom violence, utilisé au pluriel, dénote le sens d'acte de violence. Lorsque le mot est utilisé au singulier, il contribue à conceptualiser les faits de violence, afin d'en proposer une représentation générale et abstraite et d'étudier leur cause.

Le nom violence est également utilisé pour nommer des problèmes publics, c'est-à-dire des faits sociaux qui revêtent un caractère problématique et qui font intervenir l'État (Neveu 1999). Des syntagmes nominaux comme « violences faites aux femmes » et « violences policières », qu'on trouve fréquemment dans le discours médiatique, renvoient ainsi à des faits sociaux créés collectivement qui nécessitent d'avoir un nom pour exister et être représentés (Searle 1998 [1995]). Cependant, la dénomination des problèmes publics n'est pas toujours unanimement

acceptée dans la société., cet acte de nommer entraîne des débats et fait intervenir de nouveaux discours sur ces désignations.

C'est dans ce cadre que l'on va s'interroger sur la façon dont le nom violence contribue à l'émergence de phénomènes sociaux. On se posera également la question de savoir quelles sont les caractéristiques de sa manifestation lorsque violence renvoie à un fait social créé par la volonté humaine. Pour répondre à ces questions, on se basera sur l'hypothèse selon laquelle le mot acquiert de nouveaux sens en discours dans l'établissement des faits sociaux, et que son usage, lorsqu'il s'y réfère, est différent de celui qui est attendu au sein du système de la langue. Pour entreprendre l'analyse, l'approche de la sémantique discursive (Lecolle et al.

2018), qui considère le sens d'un mot comme étant construit en et par le discours a été retenue. L'étude présentée portera plus particulièrement sur la nomination (Siblot 2001) des problèmes publics et sur celle des faits s'y rapportent. L'objectif est d'observer la production du sens en prenant en compte la dimension du discours et du réel.

L'analyse est réalisée sur un corpus constitué d'articles de presse parus dans Le Monde, Le Figaro et Libération pendant les périodes où il y a eu une production discursive abondante sur ces sujets dans les médias. Le corpus ainsi construit est traité à l'aide de l'outil de textométrie TXM (Heiden et al. 2010) afin d'analyser les particularités statistiques du mot violence qui contribue à la construction des problèmes publics par sa fonction descriptive des faits qui les constituent et par sa fonction dénominative.

Cette analyse devrait contribuer à une meilleure compréhension des phénomènes sociaux qui émergent dans le discours médiatique grâce à la mise au jour des usages d'un nom abstrait dont le sens et le référent fluctuent en discours.

Bibliographie

Dewey, J. (2010 [1927]). Le public et ses problèmes (traduit par J. Zask), Paris, Gallimard.

Heiden, S., Mague, 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 – JADT 2010, 1021-1031.

Lecolle, M., Veniard, M., Guerin, O. (2018). Pour une sémantique discursive : propositions et illustrations, Langages, n°201, 35-54. https://doi.org/10.3917/lang.210.0035

Neveu, E. (1999). L'approche constructiviste des « problèmes publics ». Un aperçu des travaux anglo-saxons. Études de communication, n°22, 41-58. https://doi.org/10.4000/edc.2342

Searle, J. (1998 [1995]). La construction de la réalité sociale (traduit par C. Tiercelin), Paris, Gallimard.

Siblot, P. (2001) : De la dénomination à la nomination. Les dynamiques de la signifiance nominale et le propre du nom. Cahiers de praxématique, n°36, p. 189-214. https://doi.org/10.4000/praxematique.368

Les opérations de reformulation dans des revues de vulgarisation scientifique françaises et chinoises : premiers éléments d'analyse

LI Xuemin

Université Paris Nanterre - MoDyCo

xuemin.li@parisnanterre.fr

Mots-Clés : analyse du discours contrastive, vulgarisation scientifique, reformulation

Le discours de vulgarisation scientifique (DVS) est envisagé comme un type particulier de discours « rapportant » et « traduisant » du discours spécialisé (Mortureux, 1988 : 119, 142). Il permet donc de transmettre des connaissances du monde scientifique vers le grand public sans pour autant exiger de ce dernier qu'il dispose d'une base de connaissances spécifiques. C'est à ce niveau qu'entrent en jeu des opérations de reformulation auxquelles recourt fréquemment le DVS. Ils mettent en place des réécritures, des équivalences sémantiques, et des désignations coréférentielles autour des termes scientifiques (Reboul-Touré, 2021). Ces différentes opérations de reformulation facilitent en effet la compréhension des termes scientifiques et des notions abstraites, tout en favorisant la diffusion des savoirs au-delà des cercles restreints de spécialistes (Mortureux & Petit, 1989 : 43).

Cette communication s'inscrit dans le champ de l'analyse du discours contrastive (Claudel *et al*., 2013 ; von Münchow, 2021). La similarité des définitions entre le terme français « reformulation » et le terme chinois « chongshu » (Shen, 2009 ; Li, 2020), ainsi que la récurrence des opérations de reformulation dans les DVS des deux langues, mettent en lumière la possibilité de considérer le phénomène de reformulation comme un « concept comparatif » (Haspelmath, 2009 : 26). Ce concept joue un rôle de point de passage. Il permet une étude comparative entre les deux langues, en dépit des divergences de catégories linguistiques, d'ancrage culturel et de modes respectifs de linéarisation de l'information entre les deux langues à l'étude (Claudel, 2004 : 27). À ce stade, cette étude se concentre sur des formes des opérations de reformulation dans les DVS en français et en chinois qui portent notamment sur des sujets scientifiques d'actualité tels que les récentes avancées en intelligence artificielle, les découvertes liées à l'exploration spatiale, et les innovations en biotechnologie.

Afin de garantir la représentativité et la comparabilité, les corpus de cette étude se composent d'articles publiés tout au long de l'année 2022, sélectionnés parmi des magazines de vulgarisation scientifique à grand tirage dans les deux langues, destinés à un public non

spécialisé en sciences. Ces magazines abordent un large éventail de sujets, allant de la biologie à la médecine, l'astronomie, la géographie, l'environnement, tout en suivant de près les innovations technologiques et les découvertes récentes. Du côté chinois, nous avons retenu les magazines comme *Baike Zhishi*[1] et *Kexue 24 xiaoshi*[2] ; du côté français, nous avons sélectionné *Science & Vie*, et *Sciences et Avenir*.

Dans cette communication, nous commencerons par définir le concept de « reformulation » en français et son équivalent en chinois, « chongshu ». Ensuite, nous mènerons une réflexion théorique sur ce concept en nous inscrivant dans le domaine du DVS.

Enfin, nous proposerons une ébauche d'analyse comparative des formes de reformulation identifiées dans le corpus des deux langues en explorant les relations lexicales et sémantiques entre le reformulé (terme scientifique) et le reformulant (terme explicatif), qui contribuent à la construction du « paradigme désignationnel » (Mortureux & Petit, 1989 : 45).

Cette étude pourrait apporter une perspective novatrice, étant donné que les recherches menées en Chine sur la reformulation se sont jusqu'à présent essentiellement orientées vers l'enseignement et l'apprentissage de l'anglais dans le contexte de la didactique des langues et des cultures, ainsi que sur les comparaisons entre le chinois et l'anglais (Shen 2009 : 8).

[1] Baike Zhishi : 百科知识 en caractères chinois, cela signifie « Connaissances de l'encyclopédie »

[2] Kexue 24 xiaoshi : 科学 24 小时 en caractères chinois, cela signifie « science en 24 heures »

Bibliographie

Claudel, Ch. (2004). La notion de figure : propositions méthodologiques pour une approche comparée du genre interview de presse en français et en japonais. *Travaux Neuchâtelois de Linguistique*, 40, 27-45. <https://www.revue-tranel.ch/article/view/2597/2301>

Claudel, Ch., Von Münchow, P., Pordeus Ribeiro M., Pugnière-Saavedra, F., Tréguer-Felten, G. (dir.), (2013). *Cultures, discours, langues*. Nouveaux abordage, Limoges, Lambert-Lucas. <http://www.lambert-lucas.com/wp-content/uploads/2020/02/culture_discours_langues_oa.pdf>

Haspelmath, M. (2009). Pourquoi la typologie des langues est-elle possible ? *Bulletin de la Société Linguistique de Paris,* 104(1), 17-38.

Li, X. (2020). *Xiandai hanyu huanyan biaoji goushi yanjiu [Étude des marqueurs de reformulation dans la langue chinoise moderne]*, Thèse de doctorat en linguistique appliquée, Shanghai shifan daxue [Université normale de Shanghai]. <https://www.cnki.net/KCMS/detail/detail.aspx?dbcode=CDFD&dbname=CDFDLAST2020&filename=1020732966.nh&uniplatform=OVERSEA&v=zdrCwxuKYFTyeTq_0fG36_Nmtvni3GVNdJFcJtcI177cd2OTCkJoNw0bgbQjoj4r>

Mortureux, M.-F. (1988). La vulgarisation scientifique, parole médiane ou dédoublée ? In Jacobi, D. & Schiele, B (dir.), *Vulgariser la science le procès de l'ignorance* (p. 119148). Seyssel, Champ Vallon.

Mortureux, M.-F., & Petit, G. (1989). Fonctionnement du vocabulaire dans la vulgarisation et problèmes de lexique. *Documentation et recherche en linguistique allemande contemporain*, Vincennes, 40(1), 41-62. <https://doi.org/10.3406/drlav.1989.1076>

von Münchow, P. (2021). *L'Analyse du discours contrastive*. Théorie, méthodologie, pratique, Limoges, Lambert-Lucas.

Reboul-Touré, S. (2021). La reformulation s'adapte-t-elle aux genres de la vulgarisation scientifique ? In Ablai, D., Gonçalves, M. & Silva, F. (éds.), Reformuler. *Une question de genres ?* (p. 165-184). Húmus.
<https://run.unl.pt/bitstream/10362/141087/1/9789897557156_Reformular_uma_questao_generos_DIGITAL_1_1_.pdf>

Shen, P. (2009). *Xiandai hanyu pianzhang zhong de chongshu yanjiu [la recherche sur les reformulations dans des discours chinois contemporains]*, Thèse de doctorat en linguistique appliquée, Huadong shifan daxue [Université de l'est de la Chine].

<https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2009&filename=2009187094.nh>

Les voix derrière le masque : comparaison d'approches d'analyse de la subjectivité dans le discours de presse francophone

ESCOUFLAIRE Louis

Institut Langage & Communication (ILC), UCLouvain, Belgique

louis.escouflaire@uclouvain.be

Mots-Clés : discours journalistique ; subjectivité ; apprentissage automatique ; annotation

L'objectivité est depuis longtemps une notion fondamentale dans le monde journalistique occidental, mais sa réalisation reste un défi qui a toujours animé des discussions passionnées (Schudson, 2001). De nombreux journalistes s'efforcent d'écrire des articles aussi impartiaux et neutres que possible, dans l'espoir de transmettre aux lecteurs des faits bruts sans être influencés par leurs biais personnels (Wallace, 2020). Cependant, la nature même du processus journalistique font de l'objectivité un idéal inatteignable (Steensen, 2017). Pour atténuer la subjectivité inhérente au processus de création de nouvelles, les journalistes utilisent plusieurs techniques, conformément à ce que Tuchman (1972) appelle le « rituel stratégique de l'objectivité ». Celui-ci se réalise à travers une série de mécanismes de neutralisation conçus pour masquer les opinions personnelles des journalistes dans le contenu des textes qu'ils écrivent (Koren, 2004).

Dans l'ère du numérique, comprendre comment mesurer à quel point les opinions de journalistes ou de médias influencent le contenu d'un article de presse est une problématique de plus en plus importante (Levy, 2021). L'émergence du web et des réseaux sociaux comme canaux de diffusion majoritaires de l'information a rendu la distinction entre 'information' et 'opinion' de plus en plus floue pour les utilisateurs (Ianucci & Adair, 2017). Dans ce contexte, la présence parfois implicite de l'opinion des auteurs dans les articles de presse impacte non seulement la crédibilité et la fiabilité des sources d'information, mais ont également des implications profondes pour la littératie médiatique, façonnant en particulier la manière dont les plus jeunes lecteurs et lectrices interprètent et interagissent avec les informations qu'ils et elles rencontrent (Ku et al., 2019).

S'appuyant sur les travaux de Benveniste (1966), Kerbrat-Orecchioni a été la première à chercher à faire l'inventaire des traces énonciatives qui peuvent apparaître dans le discours en français (1970). Elle définit les unités subjectives comme « les procédés linguistiques par lesquels le locuteur imprime sa marque à l'énoncé, s'inscrit dans le message (implicitement ou

explicitement) et se situe par rapport à lui ». Cette décomposition de la subjectivité linguistique en unités discursives a été reprise dans divers travaux cherchant à analyser la présence du locuteur dans différentes formes de discours (Ho-Dac & Küppers, 2011 ; Chaput, 2019). Plus tard, Wiebe et al. (2004) affirme qu'« un discours ne peut être qualifié de discours objectif que s'il ne contient aucun indicateur significatif de subjectivité ». Les travaux de Janyce Wiebe consistant à automatiser l'analyse de la subjectivité dans le discours ont inspiré plusieurs chercheurs en traitement automatique du langage à améliorer les méthodes destinées à cette tâche cruciale (Krüger et al., 2017 ; Alhindi et al., 2020).

Notre projet de recherche vise à mettre en lumière les mécanismes linguistiques de la subjectivité dans le discours de presse en français. Notre travail s'intéresse uniquement au niveau textuel, bien que nous soyons conscients que le discours journalistique possède une dimension hautement extratextuelle relative aux choix éditoriaux des journalistes et des médias (Charaudeau, 2006 ; Steensen, 2017).

Nous croisons dans notre travail plusieurs méthodologies et développons une approche qui nous permet de confronter des analyses linguistiques plus traditionnelles (à travers une expérience d'annotation et des entretiens semi-directifs) avec des méthodes issues du traitement automatique du langage (par l'utilisation de modèles statistiques et de grands modèles de langage).

D'abord, trente-six étudiants et étudiantes en journalisme ont été amenés à évaluer la subjectivité de 150 extraits d'articles de presse belge et de surligner dans ces textes les mots qu'ils envisageaient comme des indicateurs de la subjectivité de l'auteur, dans le cadre d'une expérience d'annotation étalée sur quatre semaines. En parallèle, seize entretiens avec des journalistes québécois et belges francophones ont été réalisés, au sujet des pratiques rédactionnelles utilisées et enseignées aux journalistes pour neutraliser leurs opinions personnelles dans leurs productions écrites. Ces deux expériences nous permettent d'identifier quels éléments textuels sont considérés, par ses récepteurs (lecteurs) et/ou par ses producteurs (journalistes), comme des marqueurs potentiels de subjectivité dans le discours de presse.

Ensuite, nous avons constitué un corpus de 80 000 articles publiés par huit médias francophones belges et québécois et classés par les médias comme des articles appartenant aux genres journalistiques de l'information ou de l'opinion. A l'aide de classifieurs statistiques, nous extrayons de l'état de l'art sur la subjectivité linguistique les traits textuels les plus significatifs pour la classification d'articles d'information et d'opinion. Dix-huit indicateurs de subjectivité ressortent, selon différentes catégories : morphosyntaxiques, lexicaux et stylistiques. En parallèle, nous peaufinons (fine-tune) sur le corpus de 80 000 articles le grand

modèle de langage CamemBERT (Martin et al., 2019), pré-entraîné sur des données en français, pour la classification d'articles d'opinion et d'information. En utilisant des méthodes d'explicabilité basées sur l'attention pour déterminer quels éléments textuels ont le plus d'influence sur les décisions du modèle (Chefer et al., 2021), nous extrayons (à la suite d'une analyse qualitative des explications produites) de nouveaux indicateurs de subjectivité textuelle utilisés par CamemBERT dans la classification. Le modèle basé sur des traits atteint une précision de classification moyenne de 88.8%, tandis que le CamemBERT fine-tuné obtient 96.3%. Cependant, nos résultats permettent de mettre en perspective l'importance de l'explicabilité des modèles de classification dans une tâche comme la nôtre, pour laquelle les décisions du modèle choisi peuvent avoir d'importantes implications éthiques.

Les observations obtenues au cours de ces différentes expériences sont enfin confrontées, et nous examinons les indices sur lesquels se basent les lecteurs et auteurs humains pour déterminer si les articles de presse sont plus ou moins influencés par la subjectivité des journalistes, ainsi que les traits textuels utilisés par des modèles statistiques et des grands modèles neuronaux pour classer les mêmes articles. Bien qu'une série d'indicateurs de subjectivité identifiés dans les analyses automatiques recoupent ceux mentionnés par les lecteurs et les journalistes dans nos analyses qualitatives (comme la présence de déictiques, de modalisateurs, et la complexité du texte), chaque approche fait également ressortir des marqueurs exclusifs (présence du pronom on, niveau d'abstraction lexicale et de fréquence subjective).

Une limitation de ce projet concerne les données utilisées, qui regroupent uniquement des médias belges et québécois. Cependant, ces paysages journalistiques ont rarement été étudiés, et leur hybridité linguistique et géographique en font des sujets d'étude particulièrement intéressants dans une visée constrastive. Il est également important de souligner que les étiquettes information et opinion attribuées aux articles de notre corpus sont basées sur les catégories dans lesquelles ces articles sont déposés sur les sites web des médias dont nous les avons extraits, et dépendent donc potentiellement de choix éditoriaux dont il convient de tenir compte dans nos interprétations. Nos conclusions contribuent néanmoins à une meilleure compréhension de la façon dont la subjectivité est construite et perçue au niveau du texte dans le discours de presse en français.

Bibliographie

Alhindi T., Muresan S., & Preotiuc-Pietro D. (2020). Fact vs. Opinion: The Role of Argumentation Features in News Classification. *Proceedings of the 28th International Conference on Computational Linguistics,* 6139-6149. DOI : 10.18653/v1/2020.coling-main.540.

Benveniste E. (1966). De la subjectivité dans le langage. *Problèmes de linguistique générale*, Paris, Gallimard (coll. Bibliothèque des sciences humaines), 258-266.

Chaput L. (2019). Sur quelques marques de subjectivité dans le journalisme d'information politique de 1945 à 2015 au Québec. *Mots, 119*, 151-168. DOI : 10.4000/mots.24586.

Charaudeau, P. (2006). Discours journalistique et positionnements énonciatifs. Frontières et dérives. Semen. *Revue de sémio-linguistique des textes et discours, 22.*

Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 782-791.

Ho-Dac L.-M., & Küppers A. (2011). La subjectivité à travers les médias : Étude comparée des médias participatifs et de la presse traditionnelle. *Corpus, 10*, 179-199.

Iannucci, R., Adair, B. (2017). *News or opinion? Online, it's hard to tell.* www.poynter.org/ethicstrust/2017/news-or-opinion-online-its-hard-to-tell/ (last accessed 24/05/2023).

Kerbrat-Orecchioni C. (2009). *L'énonciation : de la subjectivité dans le langage*. Armand Colin.

Koren, R. (2004). Argumentation, enjeux et pratique de l'« engagement neutre »: le cas de l'écriture de presse. *Semen. Revue de sémio-linguistique des textes et discours, 17*.

Krüger K. R., Lukowiak A., Sonntag J., Warzecha S., & Stede M. (2017). Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, *23(5),* 687-707. DOI : 10.1017/S1351324917000043.

Ku, K. Y., Kong, Q., Song, Y., Deng, L., Kang, Y., & Hu, A. (2019). What predicts adolescents' critical thinking about real-life news? The roles of social media news consumption and news media literacy. *Thinking Skills and Creativity, 33*, 100570.

Levy, R. E. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review, 111(3)*, 831-870.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D. & Sagot, B. (2019). Camembert: a tasty french language model. *arXiv:1911.03894.*

Schudson, M. (2001). The objectivity norm in American journalism. *Journalism, 2(2),* 149-170.

Steensen S. (2017). Subjectivity as a Journalistic Ideal. *Putting a Face on it: Individual Exposure and Subjectivity in Journalism*. Cappelen Damm Akademisk, 25-47.

Tuchman G. (1972). Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of Sociology, 77(4),* 660-679.

Wallace, L. R. (2020). *The view from somewhere: undoing the myth of journalistic objectivity.* University of Chicago Press.

Wiebe J., Wilson T., Bruce R., Bell M. & Martin M. (2004). Learning Subjective Language. *Computational Linguistics, 30(3),* 277-308. DOI : 10.1162/0891201041850885.

Session 2

Discours dans l’espace numérique

Expression de l'évaluation dans le commentaire sur Internet

OSHCHEPKOVA Liudmila

Université de Lorraine

oschepkova.liudmila@gmail.com

Mots-Clés : commentaire sur Internet ; genres numériques ; évaluation positive et négative ; caractéristiques axiologiques ; moyens d'expression de l'évaluation.

Il est difficile d'imaginer notre vie sans les réseaux sociaux. La communication que nous y entretenons au moyen de tchats, de publications, de messages, de commentaires, etc. est déjà comparable à la communication face à face. Plusieurs chercheurs se sont tournés vers l'analyse de genres numériques tels que les réseaux sociaux (M.-A. Paveau, 2013), les blogs (A. M. Gjesdal, Ø. Gjerstad, 2014), le courrier électronique (E. Gajewska, 2016), les tweets (M.-A. Paveau, 2013) et bien d'autres.

Notre présentation propose une analyse du genre " le commentaire sur Internet " en termes de caractéristiques axiologiques. Un commentaire sur Internet est un message exprimant l'opinion de l'utilisateur sur une publication sur Internet ou commentant le contenu d'un autre commentaire (E. Yu. Viktorova, K. V. Panteeva, 2023 : 67). En raison de la multidimensionnalité des genres sur Internet, ceux-ci sont fortement hiérarchisés : il existe les hypergenres, les genres, les sous-genres et les genres associés (M. Bonhomme, 2015 : 32). Le commentaire sur Internet est un sous-genre (genre non autonome), qui accompagne le texte original (commenté) (L. Yu. Shchipitsina, 2015 : 529). Il ne peut donc pas être étudié indépendamment du genre principal, la publication elle-même. La particularité du commentaire sur Internet réside dans la combinaison des caractéristiques du discours oral et du discours écrit. Comme le précise C. Combe (2019 : 51), " à l'heure de la numérisation généralisée de la société, après les genres écrits et les genres oraux, ce sont désormais les « genres numérique » ". Du format oral de la communication, le commentaire sur Internet a hérité de l'informalité, de la concision et d'un haut degré d'émotivité, tandis que du format écrit, il a emprunté la structuration du discours et l'utilisation des possibilités graphiques d'Internet.

Souvent, l'expression de l'évaluation dans un commentaire devient plus importante que la transmission d'un message sémantique. Selon P. Pupier (1998 : 53), il faut moins d'informations pour reconnaître qu'une unité lexicale représente quelque chose de " bon " ou de " mauvais " que pour savoir ce qu'elle signifie. Par conséquent, dans les commentaires sur

Internet concernant des publications sur des sujets sensibles, la caractéristique évaluative est lue en premier, ensuite seulement – la caractéristique sémantique.

Notre étude porte sur l'élément évaluatif dans les commentaires du réseau social " Instagram ", en utilisant l'approche discursive. Avec le début du conflit militaire russoukrainien, de nombreuses personnalités publiques en Russie ont exprimé leur position antiguerre sur Internet dès le lendemain, notamment sous la forme d'une publication sur Instagram. Nous avons sélectionné des exemples de publications de différentes célébrités (chanteurs, humoristes, présentateurs de télévision, psychologues, sportifs, etc.) de nationalité russe qui se sont exprimées contre la guerre et analysé les commentaires de ces publications.

Chaque jugement évaluatif " met en jeu a minima des interactions entre une cible évaluée, une source évaluatrice et un contexte d'évaluation " (A. Jackiewicz, 2014 : 5). Dans le cadre de notre étude, le sujet ou la source de l'évaluation sont les utilisateurs d'Instagram. L'objet principal de l'évaluation est, en général, le contenu anti-guerre de la publication ou la personnalité médiatique elle-même qui peut être évaluée sur la base de ses qualités personnelles, de ses activités professionnelles, de sa vie privée, etc.

Le contexte d'évaluation joue un rôle important, car il peut non seulement renforcer ou affaiblir les caractéristiques axiologiques, mais aussi les remplacer par des caractéristiques opposées. Dans notre étude, le contexte est le conflit militaire russo-ukrainien, sujet dont la discussion est impossible sans jugements évaluatifs. Le milieu de communication est le réseau social, qui entraîne le choix des moyens d'expression de l'évaluation.

La composante évaluative des commentaires sur l'internet se caractérise par une riche palette de moyens d'expression à différents niveaux de langue, qui comprennent des moyens graphiques, morphémiques, lexicaux, phraséologiques et syntaxiques.

Parmi les moyens graphiques d'expression de l'évaluation figurent l'utilisation de lettres majuscules ou minuscules en contradiction avec les règles d'orthographe, la multiplication des lettres, l'utilisation d'émoticônes.

Une autre façon d'exprimer l'évaluation est d'utiliser des moyens morphémiques : **quasi-**psychologue, patriot**ounet**s (en russe, le sens est exprimé par le suffixe diminutif –ик : патриот**ик**и), **pseudo**-patriotisme.

Le plus souvent, l'évaluation est exprimée par des moyens lexicaux. Par exemple, les internautes utilisent un vocabulaire évaluatif avec la signification du caractère et des qualités personnelles pour évaluer positivement l'auteur d'une publication (personnalité médiatique) : *courage, bravoure, vaillance, volonté, honnêteté, humanité, vérité* (*Courage et honnêteté dans*

vos paroles, Lioudmila !). Les expressions métaphoriques ne sont pas rares (*un grand homme, un vrai homme*), tout comme le vocabulaire somatique *: une personne au grand cœur*.

Les moyens phraséologiques d'exprimer l'évaluation comprennent l'utilisation d'expressions idiomatiques : *A soutenu l'Ukraine* ***et s'est mis sur la paille*** *; Vous semblez* ***vivre les yeux fermés*** *; Faites de la psychologie et* ***ne fourrez pas votre nez dans*** *la politique, la guerre, etc*.

L'évaluation peut également être exprimée par des moyens syntaxiques, par exemple à l'aide de questions rhétoriques *(Pour vous, le monde ne s'est écroulé que maintenant ????? Ou ne nous préoccupons-nous que de ce qui nous a touchés ?*) et des appels. Les utilisateurs qui soutiennent une célébrité s'adressent à elle par son prénom, son prénom et son patronyme (forme respectueuse en russe), son prénom et son nom, en utilisant le mot *honoré*, ainsi que des diminutifs affectueux. Les marqueurs d'une évaluation négative sont l'adresse simplement par le nom de famille, les diminutifs avec une connotation méprisante et l'utilisation sarcastique de la forme polie officielle *господин* (monsieur).

Tous ces moyens d'expression de l'évaluation peuvent être utilisés de manière complexe, par exemple pour créer un effet sarcastique.

(1) *Où t'enfuiras-tu ensuite, Maximka ?*))))) Dans cet exemple, le sujet du sarcasme est la décision de quitter la Russie. Une appellation par le prénom est utilisée. *Maksimka* est un diminutif du prénom *Maxime*, formé à l'aide du suffixe *-k*. En outre, le verbe neutre *partir* est remplacé par le verbe *s'enfuir*, qui a une connotation négative. Le sarcasme est souligné par l'utilisation de parenthèses comme émoticônes. Le commentaire concerne l'humoriste Maxime Galkine, parti en Israël après le début du conflit russo-ukrainien. Ce commentaire a été laissé après le début du conflit militaire entre Israël et la Palestine. L'auteur du commentaire ironise donc sur le fait que l'humoriste va devoir émigrer une fois de plus.

(2) *Prends la vieille et cours au bureau de recrutement)))) ou sa date de péremption est dépassée, elle risque de ne pas y arriver ?* L'auteur de ce commentaire se moque du fait que Maxime Galkine est marié à la chanteuse Alla Pougatcheva, qui a 27 ans de plus que lui. Ce commentaire est basé sur l'image négative d'un produit gâté, auquel l'auteur compare Alla Pougatcheva. La chanteuse elle-même est appelée *la vieille*, un lexème à connotation négative (en russe *старуха*). Le sarcasme est également souligné par l'utilisation d'un émoticône.

Ainsi, la composante axiologique est primordiale pour le genre " le commentaire sur Internet ", car il est important pour les utilisateurs de partager leur opinion et d'évaluer le message et son auteur. Les commentaires positifs sont généralement beaucoup plus courts et se limitent à des remerciements et à une brève description favorable de la célébrité ou du

contenu du message. Les commentaires négatifs, en revanche, sont plus longs, variée, émotionnels et plus créative. Cette différence de longueur peut être expliquée par la notion de " preference for agreement " en analyse conversationnelle, qui suggère que l'accord est généralement formulé de manière plus courte que le désaccord. De plus, l'anonymat et la distance permettent d'exprimer l'évaluation négative de manière plus explicite et détaillé, ce qui est gêné par les normes de l'étiquette dans la communication réelle.

Bibliographie

[1] Bonhomme M. (2015) La problématique des genres de discours dans la communication sur Internet. *Travaux neuchâtelois de linguistique*, 63, 31-47.

[2] Combe C. (2019) Les genres numériques de la relation. *Langage et société*, 2019/2 (N° 167), 51-80. DOI : 10.3917/ls.167.0051. URL : https://www.cairn.info/revuelangage-et-societe-2019-2-page-51.htm

[3] Gjesdal A. M., Gjerstad Ø. (2014) Web 2.0 et genres discursifs : l'exemple de blogs sur le changement du climat. *Synergies Pays Scandinaves*, n° 9, 49-61.

[4] Jackiewicz A. (2014) Études sur l'évaluation axiologique : présentation. *Langue française*, 2014/4 n° 184, 5-16. DOI : 10.3917/lf.184.0005. URL : https://shs.cairn.info/revue-langue-francaise-2014-4-page-5?lang=fr.

[5] Paveau M.-A. (2013) Genre de discours et technologie discursive. Tweet, twittécriture et twittérature. *Pratiques*, 157-158. URL : http://journals.openedition.org/pratiques/3533 ; DOI : https://doi.org/10.4000/pratiques.3533

[6] Paveau M.-A. (2013) Analyse discursive des réseaux sociaux numériques. Dictionnaire d'analyse du discours numérique. *Technologies discursives* [Carnet de recherche], http://technodiscours.hypotheses.org/?p=431

[7] Pupier P. (1998). Une première systématique des évaluatifs en français. *Revue québécoise de linguistique*, 26 (1), 51-78. https://doi.org/10.7202/603144ar

[8] Shchipitsina L. Yu. (2015) Genre status of an online comment. *Vestnik Baškirskogo universiteta*, № 2. URL: https://cyberleninka.ru/article/n/zhanrovyy-status-setevogokommentariya

[9] Viktorova E. Yu., Panteeva K. V. (2023) An Internet comment as a speech genre: Axiological aspect. *Speech Genres*, vol. 18, no. 1 (37), 66-73. https://doi.org/10.18500/2311-0740-2023-18-1-37-66-73, EDN: HGVVCL

Analyse du genre de discours : le cas des sites web d'agences de voyages francophones

PHAM Tran Hanh Trang, CARDON Rémi, WILKENS Rodrigo, FRANÇOIS Thomas

Université Catholique de Louvain,

tran.pham@uclouvain.be, remi.cardon@uclouvain.be, rodrigo.wilkens@uclouvain.be, thomas.francois@uclouvain.be

Mots-Clés : Discours touristique numérique ; genre de discours ; structure des tours

L'essor des voyages internationaux et de l'utilisation d'Internet pour organiser les voyages a rendu cruciale la communication via les sites web touristiques, notamment pour les agences de voyage (Constantinides *et al*., 2010, Gémar, 2014). Les caractéristiques de ces sites sont étroitement liées aux stratégies commerciales des agences et influencent les décisions et la satisfaction des touristes (Violino, 2001). Dans les études portant sur le discours touristique en ligne, seule une minorité intègre la notion de genre de discours (Malenkina et Ivanov, 2018). Le genre est défini comme une classe d'événements communicatifs dont les membres partagent un ensemble d'objectifs communicatifs. Ces objectifs sont reconnus par les membres experts de la communauté discursive parentale et constituent ainsi la raison d'être du genre. Cette raison d'être façonne la structure schématique du discours et influence et contraint le choix du contenu et du style (Swales, 1990, p. 58). Comprendre le genre et ses éléments essentiels dans le tourisme numérique permet aux agences de voyage de rédiger des textes efficaces et d'accomplir les trois fonctions de communication typiques : attirer l'attention, informer et persuader (Bhatia, 2004). Cette approche offre un cadre pour analyser et améliorer la communication sur les sites web touristiques.

Les chercheurs ont abordé divers genres de discours touristique en ligne, notamment des sites web des offices de tourisme (Cheong, 2013), des destinations touristiques spécifiques (Huang, 2015 ; Hui *et al*., 2020 ; Nguyen et Modehiran, 2023), des guides touristiques (Alali *et al*., 2019), ou des brochures (Öztürk et Şafak, 2014). Cependant, les sites web touristiques des agences de voyage sont peu analysés. En outre, les caractéristiques linguistiques analysées dans des études précédentes sont généralement limitées (au niveau morphosyntaxique, syntaxique ou lexical). De plus, peu de recherches ont été enregistrées sur les hyperliens – caractéristique importante pour le genre web.

Ainsi, notre étude consiste à analyser 80 pages d'accueil des sites web d'agences de voyage francophones afin de déterminer s'il s'agit d'un genre de discours à part entière et à le caractériser. Pour ce faire, nous utilisons le corpus du français numérique *frWac* (Baroni *et al*., 2009) et un corpus de sites web touristiques collectés spécifiquement pour cette étude. Nous comparons d'une part les discours touristiques numériques à un échantillon de discours numériques variés, et d'autre part à comparer les discours des sites d'agence de voyage à 90 pages d'accueil des autres discours numériques touristiques (les blogs, les sites d'offices de tourisme et les sites de points d'intérêt touristiques). Pour ces comparaisons, nous exploitons l'analyse rhétorique via la structure des « tours » (*moves* en anglais), développée par Askehave et Nielsen (2005) pour l'analyse du genre des sites web. Parallèlement, nous mobilisons FABRA (Wilken *et al*., 2022), un outil de traitement automatique des langues, pour l'extraction d'une série de caractéristiques linguistiques et discursives, et effectuons également une analyse des caractéristiques numériques via des hyperliens.

Les résultats montrent que les sites web des agences de voyage constituent un genre de discours en soi, dont les caractéristiques linguistiques et numériques diffèrent bien d'autres discours touristiques numériques. Tout d'abord, les agences de voyage se démarquent par l'usage d'un langage riche, diversifié et sophistiqué, qui semble moins accessible au grand public. Ensuite, elles accordent une grande importance à leur image et cherchent à impliquer activement les internautes dans la communication. De plus, il semble qu'elles ne cherchent pas spécifiquement à créer des communautés de discours. La création d'une communauté de discours, selon Askehave et Nielsen (2005), permet aux utilisateurs fidèles ou fréquents d'établir des communautés au sein du site web (par exemple grâce à la fonction de connexion au site web).

Les contributions de cette étude sont multiples. Notre première contribution consiste à éclairer la problématique de la catégorisation d'un genre, en l'occurrence les sites web d'agences de voyage, à l'aide d'une approche empirique et à large spectre (au niveau des caractéristiques textuelles considérées). En effet, cette étude se base sur un jeu étendu de caractéristiques linguistiques, grâce à des outils de traitement automatique du langage (TAL), qui permettent de dépasser les « classiques » étiquettes morpho-syntaxiques. Une seconde contribution concerne la prise en compte de caractéristiques typiques des pratiques numériques sur le web, au travers d'une analyse des hyperliens. Troisièmement, notre analyse dépasse le niveau textuel et exploite également la notion de « tours » pour mieux caractériser l'association entre les fonctions communicatives et la structure rhétorique des discours étudiés. Cette

tentative d'appliquer une approche fréquemment utilisée pour examiner le langage en anglais, à l'analyse du genre de discours promotionnel en français, représente une ultime contribution.

Bibliographie

[1] Alali, B. A., Ali, A. M., et Ali, A. M. (2019). Genre-based Analysis of Travel Guides: A Study on Malaysia, Thailand and the Philippines. *LSP International Journal*, 6(2).

[2] Askehave, I., et Nielsen, A. E. (2005). *Digital Genres: A Challenge to Traditional Genre Theory*. Information Technology and People, 18(2), 120-141. https://doi.org/10.1108/09593840510601504

[3] Baroni, M., Bernardini, S., Ferraresi, A., et Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *LRE*, 43.3: 209-226.

[4] Bhatia, VK. (2004). *Worlds of Written Discourse : A Genre-Based View*. London : Continuum.

[5] Cheong, Y. M. (2013). *A multi-dimensional genre analysis of tourism homepages and webmediated advertorials.* Doctoral dissertation, Universiti Malaya.

[6] Constantinides, E., Lorenzo-Romero, C. et Gómez, M.A. (2010). Effects of web experience on consumer choice: a multicultural approach. *Internet Research, Vol. 20 No. 2*, 188-209.

[7] Gémar, G. (2014). Influence of cultural distance on the internationalization of Spanish hotel companies. *Tourism & Management Studies, 10 (1)*, 31-36.

[8] Huang, S. (2015). A genre-based analysis of brief tourist information texts. *2015 Joint International Social Science, Education, Language, Management and Business Conference*, 191-202.

[9] HUI, W., Santhi, N., et Mungthaisong, S. (2020). A Genre Analysis Of Online English Tourist Attraction Promotional Texts In Lijiang Yunnan Province. วารสาร สห วิทยาการ สังคมศาสตร์ และ การ สื่อสาร, 3(1), 83-106.

[10] Malenkina, N. et Ivanov, S. (2018). A linguistic analysis of the official tourism websites of the Seventeen Spanish autonomous communities. *Journal of destination marketing & Management 9*, 204-233.

[11] Nguyen, T. D., et Modehiran, P. (2023). A Genre-Based Analysis of Vietnam Tourist Attraction Brochures with Pedagogical Purposes. *ABAC ODI Journal Vision. Action. Outcome, 11(1),* 43-55.

[12] Öztürk, B., et Şafak, Z. (2014). Genre analysis of a Turkish tourism brochure. Linguistics, Culture And Identity. *Foreign Language Education, 351*.

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

[14] Violino, B. (2001). E-business lurches abroad. *Internet Week*, le 19 mars.

Technologie et société dans le monde postpandémique : les réseaux numériques dans la construction de mémoires et d'évènements discursifs dans l'envers de l'histoire

ALMEIDA RUIZ Marco Antonio

Federal University of Goiás – UFG / FAPEG

marcoalmeidaruiz@gmail.com

Mots-Clés : Discours ; Numérique ; Changement technologiques ; Covid ; 19 ; Instagram.

Cette proposition de **communication orale** vise à examiner l'impact des réseaux sociaux dans la construction de sens autour de la pandémie de covid-19 au Brésil en 2020 et 2021. Cette maladie a causé des milliers de décès dans le pays et a donné une nouvelle signification au deuil pour de nombreuses familles qui ont dû dire adieu à leurs proches de manière soudaine et terrible face à un virus apparemment inconnu, mais totalement mortel. En conséquence de la mauvaise gestion d'un ancien dirigeant incompétent et sans scrupules, nous avons vu des scènes choquantes de corps enveloppés dans des sacs noirs et, en raison du grand nombre de décès par jour, il a fallu les réfrigérer dans des entrepôts frigorifiques installés à côté des hôpitaux, en créant une véritable scène de guerre et d'horreur qui était diffusée en boucle par la presse. Ce cadre de crise a radicalement modifié le mode de vie et transformé la société. Sur le plan social, d'une part, nous avons dû adopter de nouvelles habitudes d'hygiène personnelle, ainsi que nous isoler de nos proches par mesure de protection et de santé, un geste très difficile mais totalement nécessaire pour éviter une contamination et des décès rapides ; d'autre part, sur le plan économique, nous avons vu les Brésiliens qui ont perdu leur emploi, des magasins et plusieurs commerces ont fermé leurs portes et, par conséquent, le marché financier a été touché, augmentant la pauvreté et rendant l'accès aux aliments et produits de base pour la survie plus difficile.

En raison du déni du virus par certains représentants politiques, quelques conflits sociaux ont également émergé : d'un côté, des négationnistes qui ne croyaient pas aux mesures de protection, accusant les maires d'être responsables de l'augmentation de la pauvreté et du manque de ressources économiques ; d'autre côté, des chercheurs et des associations scientifiques défendant la vie et l'isolement social comme moyen préventif contre le répandu de la maladie dans le pays. Dans le domaine des arts, linguistique, sujet sur lequel nous nous concentrons dans ce travail, nous avons témoigné l'émergence de profils sur les réseaux

sociaux qui se moquaient des actions (ou de l'absence d'action) du gouvernement, apportant aussi plus de couleur, d'espoir et un bref soulagement face à tant de moments d'horreur dans notre histoire. Pour cela, notre objectif est d'analyser le *mémorial* virtuel @Museudoisolamento (en portugais) sur Instagram, en observant dans son émergence les formes de résignification de la mort et du deuil générées dans un contexte d'instabilité et de crise politique brésilienne. Plus précisément, en utilisant les présupposés théorique-méthodologiques de l'analyse du discours d'inspiration française, nous effectuerons un exercice d'analyse de certaines publications réalisées pendant la période la plus critique de la pandémie au Brésil afin de comprendre les gestes de résistance divulgués par ce matériel numérique.

Selon Orlandi (2002), la constitution du dire se fait par le biais d'une mémoire du dire dans laquelle sont marqués discursivement les effets de sens relativement stabilisés, découlant de préconstruits et d'autres discours déjà énoncés ; en ce qui concerne la formulation, celle-ci se produit effectivement à partir du moment où les conditions de production de ces énoncés sont directement ou indirectement liées aux circonstances de l'énonciation. Dans cette étude, la mort, qui n'était que des chiffres, devient des histoires de vie qui n'ont pas été perdues même face à l'indifférence et à l'omission des discours officiels. Cependant, la transformation de la mémoire se produit uniquement au niveau de la circulation, car il y a une mise à jour, transformant les individus et l'effet de sens comme résultat de la résignification d'une actualité et d'une mémoire fondées sur les conditions d'émergence de discours reflétant chaque époque et chaque formation sociale. Ainsi, lorsque nous pratiquons l'exercice d'analyse de la mémoire, nous découvrons quelque chose de beaucoup plus complexe, qui échappe à la dimension chronologique et psychologisante et acquiert des « sens entrecroisés de la mémoire mythique, de la mémoire sociale inscrite dans les pratiques, et de la mémoire, construite de l'historien ». (Pêcheux, 2010, p. 50). Autrement dit, « pour qu'il y ait mémoire, il faut que l'évènement ou le savoir enregistré sorte de l'indifférence, qu'il quitte le domaine de l'insignifiance ». (Pêcheux, 2010, p. 25). Dans notre cas, nous abordons la mémoire comme une actualisation d'un événement historique, c'est-à-dire comment la pandémie de covid-19 au Brésil a apporté des changements profonds dans les modes de relation entre les individus et a créé des pratiques qui réinterprètent la mort et le deuil à travers certains profils sur les réseaux sociaux, tels que le mémorial virtuel.

Ces matériels deviennent synonymes de nouveaux discours qui (re)racontent la mémoire du deuil déjà cristallisée dans la société. Penser au numérique et à certains profils qui échappent à cette régularité énonciative sur la mort et le deuil, en particulier en observant les

(re)significations que ces matérialités entraînent dans le jeu de résistance contre les discours officiels et négationnistes, nous permet de comprendre les nouveaux réseaux d'affiliations et de circulation de discours qui apparaissent dans différents lieux de circulation, composant ainsi cette formulation. Dire quelque chose dans une certaine instance discursive, c'est chercher dans l'histoire des traces de sens laissées dans les lignes du temps, revisitant des positions qui s'affrontent dans les affaires contemporaines. Le profil du Musée analysé est apparu comme une forme de résistance à l'horreur de la mort, au déni, où seules les statistiques froides étaient utilisées pour décrire le nombre de cas de décès et de contaminations. Aux vecteurs glacés diffusés pour comptabiliser les morts, nous avons observé le retour de la subjectivité des proches, des amours de quelqu'un, ramenés encore dans le champ de la discussion, représentant tout leur rôle et leur importance pour leurs familles et leur communauté, montrant qu'ils n'étaient pas juste un nombre.

Nous avons sélectionné quelques extraits du profil @Museudoisolamento sur Instagram afin d'analyser l'utilisation de l'art comme résistance à l'horreur de la mort et du deuil, singularisant la vie et lui donnant de la valeur. Notre travail n'a pas pour but de conclure ces questions problématisées avec cette « technologisation des discours », au contraire, nous voyons comment ces réseaux ont le potentiel de rompre aussi avec certains imaginaires sociaux qui présentent en grande partie les discours des « puissants » comme les « propriétaires » d'une « vérité ».

Bibliographie

Museu do Isolamento, *Instagram*, 6 out. 2020. Disponível em: https://instagram.com/museu.do.agora?igshid=MzRlODBiNWFlZA==. Acesso em: 18 abr. 2024.

ORLANDI, E. (2002). *Análise de Discurso: princípios e procedimentos*. Campinas: Pontes.

PAVEAU, M. A. (2021). *Análise do discurso digital: dicionário das formas e das práticas*. Campinas: Pontes Editores.

PÊCHEUX, M. (2008). *O discurso: estrutura ou acontecimento*. Campinas: Pontes.

PÊCHEUX, M. (2010). *Papel da memória.* Campinas: Pontes. ontenu de bibliographie (12pt)

Session 3

Corpus d’aujourd’hui

La textométrie et les corpus oraux. Apports et limites d'une approche outillée des effets discursifs du vieillissement.

DAGNIAUX Anaïs

ELLIADD – EA4661, Université de Franche-Comté

anais.dagniaux@edu.univ-fcomte.fr

Mots-Clés : vieillissement ; textométrie ; corpus oraux ; analyse du discours

Inscrite en analyse du discours (Charaudeau et Maingueneau, 2002 ; Maingueneau, 2021), notre thèse de doctorat porte sur les changements induits par les vieillissements typique et atypique – respectivement caractérisés par l'absence et la présence de troubles pathologiques liés à l'âge (Gerstenberg, 2015) – sur les capacités langagières et communicationnelles des personnes âgées. Nous nous intéressons plus particulièrement aux phénomènes qui ont été à ce jour délaissés par les approches cognitives tels que la cohérence textuelle (Charolles, 1978), la progression thématique (Combettes, 1988) ou encore l'expression de la subjectivité (Kerbrat-Orecchioni, 1980) dans le vieillissement et la maladie d'Alzheimer.

Dans cette perspective, nous constituons un corpus semi-contrôlé (Tellier, 2014) à partir d'une méthode originale de recueil de la parole associant restitution d'informations et production orale semi-spontanée. Les transcriptions de nos entretiens sont ensuite destinées à être explorées par une analyse textométrique (Heiden *et al*., 2010), associant prises de vue quantitatives et retours au contexte (Pincemin et Heiden, 2008).

Pour qui travaille sur des données orales, utiliser la textométrie est une option méthodologique qui n'est pas sans poser son lot de difficultés. En amont de toute analyse à proprement parler, l'étape de transcription des données orales implique de nombreux choix déterminant les résultats de la recherche, à commencer par le choix des conventions de transcription. Celles-ci doivent viser la poursuite des objectifs de la recherche tout en obéissant aux principes de fidélité et de lisibilité de la transcription (Blanche-Benveniste, 2010). Ultérieurement, l'étape de prétraitement des données (manipulations en vue de garantir l'interopérabilité (Badin *et al*., 2021), alignement des données, annotation linguistique) s'avère cruciale pour rendre les données exploitables et pertinentes en regard des objectifs de la recherche. En effet, certains phénomènes typiques de la langue parlée, à l'instar des disfluences et des marqueurs discursifs, ne s'inscrivent pas dans les catégories grammaticales existantes pour l'écrit et traditionnellement adoptées par les outils d'étiquetage automatique mobilisés en

textométrie, tels que *TreeTagger* (Schmid, 1994). Ainsi, comment espérer rendre compte des spécificités de l'oral en mobilisant des outils initialement conçus pour traiter des données écrites ?

Pointant les difficultés posées par les corpus oraux à la textométrie telles qu'elles apparaissent dans le cadre de notre recherche doctorale, notre communication a pour objectif de répondre à la question suivante : en quoi une analyse renouvelée du discours dans le vieillissement est-elle à la fois autorisée et limitée par l'état des outils textométriques ?

Notre présentation s'articulera autour de trois axes. Nous exposerons d'abord les objectifs de notre recherche, centrés sur des aspects discursifs ignorés dans la littérature, principalement nourrie par des approches issues de la neuropsychologie et de la psychologie cognitive. Nous aborderons également les motivations qui nous ont conduite à adopter une approche outillée par la textométrie, telles que la possibilité du retour au média implémentée dans *TXM* (Heiden *et al.,* 2010) permettant la prise en compte des aspects paraverbaux et non-verbaux des données (Pincemin *et al*., 2020).

Nous présenterons ensuite la méthodologie de constitution de notre corpus, que nous illustrerons à partir d'un échantillon de cinq entretiens (environ 10000 mots) effectués auprès de locuteurs caractérisés par un vieillissement typique. Ces entretiens sont transcrits à partir du guide de transcription de Heiden et Pincemin (2011) et des conventions de transcription utilisées dans le cadre du projet CorpAGEst (Bolly et Kairet, 2016 ; Bolly et Boutet, 2018), que nous avons adaptés à nos besoins. Alignées avec les enregistrements à l'aide de *Transcriber* (Barras *et al*., 2001), les données sont ensuite importées sur l'outil textométrique *TXM*, puis étiquetées morphosyntaxiquement et lemmatisées.

Parvenue à ce stade, nous détaillerons l'inadéquation des prétraitements automatiques (lemmatisation, annotation) effectués sur des données orales, qui nous amènent à une analyse critique des cadres théoriques qui sous-tendent l'outil d'une part, et à envisager d'autre part les apports d'une annotation linguistique fine et manuelle pour l'examen des phénomènes typiques de la langue parlée.

Bibliographie

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*. N°22. https://doi.org/10.4000/corpus.5752.

Barras, C., Geoffrois, E., Wu, Z. & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*. N°33. https://hal.science/hal-01690349/file/transcriber.pdf.

Blanche-Benveniste, C. (2010). *Approches de la langue parlée en français*. Ophrys.

Bolly, C. & Kairet, J. (2016). *CorpAGEst (2013-2015): "A corpus-based multimodal approach to the pragmatic competence of the elderly". Multimodal annotation guidelines.* https://corpagest.wordpress.com/wp-content/uploads/2016/06/corpagest_spannotation_manual_v1-3.pdf.

Bolly, C. T., & Boutet, D. (2018). The multimodal CorpAGEst corpus: keeping an eye on pragmatic competence in later life. *Corpora*. N°13, 279-317. 10.3366/cor.2018.0151.

Charaudeau, P. & Maingueneau, D. (2002). *Dictionnaire d'analyse du discours.* Seuil.

Charolles, M. (1978). Introduction aux problèmes de la cohérence des textes. *Langue française.* N°38, 7-41. https://doi.org/10.3406/lfr.1978.6117.

Combettes, B. (1988*). Pour une grammaire textuelle : la progression thématique.* De Boeck.

Gerstenberg, A. (2015). Langue et générations : enjeux linguistiques du vieillissement. Polzin-Haumann, C., & Schweickard, W. (dir.). *Manuel de linguistique française.* Walter de Gruyter, 314-333.

Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM : Une plateforme logicielle opensource pour la textométrie – conception et développement. *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*. Rome, Italie. http://halshs.archivesouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf.

Heiden, S. & Pincemin, B. (2011). *Guide de transcription d'entretien avec Transcriber pour TXM*. https://shs.hal.science/halshs-01341955.

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

Maingueneau, D. (2021). *Discours et analyse du discours*. 2e éd. Arman Colin.

Pincemin, B. & Heiden, S. (2008). Qu'est-ce que la textométrie ? Présentation. *Site du projet Textométrie*. https://txm.gitpages.huma-num.fr/textometrie/Introduction/.

Pincemin, B., Heiden, S. & Decorde, M. (2020). Textometry on Audiovisual Corpora: Experiments with TXM software. *JADT 2020 : 15th International Conference on Statistical Analysis of Textual Data JADT 2020*. Toulouse, France. https://shs.hal.science/halshs02779055v1/document.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of*

International Conference on New Methods in Language Processing. Manchester, RoyaumeUni. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Tellier, M. (2014). Quelques orientations méthodologiques pour étudier la gestuelle dans des corpus spontanés et semi-contrôlés. *Discours*. N°15. https://doi.org/10.4000/discours.8917.

Bootstrapper son corpus oral ou comment développer un corpus d'arabe parlé dans le petit Maghreb

ZIANE Rayan[1], MESSAOUDI Fatma Ben Barka[2]

1 : Centre de recherche inter-langues sur la signification en contexte, Université de Caen,
2 : École, mutations, apprentissages (EMA) CY Cergy Paris Université

Ziane.rayan4@gmail.com, fatma.messaoudi1@cyu.fr

Mots-Clés : linguistique de corpus ; oral ; ASR ; processus itératif

Les dernières années ont connu un essor des techniques d'analyse linguistique automatique basées sur l'adaptation (fine-tuning) de modèles pré-entraînés à partir de petits corpus annotés. Peng et al. (2022) ont notamment démontré l'efficacité d'une approche agile de développement de corpus arborés, consistant à enrichir progressivement un corpus d'apprentissage par des données issues d'analyses automatiques d'abord imprécises, corrigées manuellement par des experts. La transcription contribue au cycle de documentation des langues en se plaçant en interface des autres types d'analyses du plan phonétique à l'analyse du discours en passant par la morpho-syntaxe.

Dans cette communication, nous proposons de répliquer cette méthodologie pour le développement de corpus oraux. La démocratisation des techniques de reconnaissance automatique de la parole (ASR) ouvre de nouvelles perspectives en ce sens (Baevski et al., 2020; Pratap et al., 2023). Par affinage d'un modèle multilingue pré-entraîné sur des quantités colossales de données, quelques heures transcrites manuellement sont suffisantes afin d'entraîner un premier système imparfait pour faciliter la transcription manuelle de plus de données avant de parfaire le système et ses prédictions durant un processus itératif qui se nourrit des transcriptions validées par l'expert.

La détection automatique des tours de parole et la discrimination des locuteurs bénéficient également de ce bond technologique (Plaquet et Bredin, 2023). Là encore, le modèle préentraîné généraliste peut être adapté avec des segmentations éditées manuellement. Cette approche permettrait de gérer efficacement les spécificités du langage oral, telles que les chevauchements de paroles, les faux départs et les interruptions fréquentes. En identifiant précisément les tours de parole, la qualité de la transcription peut être améliorée et rendre compte fidèlement des dynamiques de l'interaction entre locuteurs.

En outre, Guillaume et al. (2022) montraient comment cette méthode ouvre des perspectives prometteuses pour le traitement et l'inclusion des langues peu dotées. Les langues minoritaires ou moins représentées dans les grands ensembles de données peuvent bénéficier de cette approche itérative.

Le contexte de notre contribution est celui du continuum linguistique de l'arabe parlé au petit Maghreb (Algérie, Maroc, Tunisie), dont les variétés orales sont dépourvues de système orthographique standardisé. Pendant longtemps, le paysage linguistique dans le petit Maghreb se caractérisait par la prédominance de la variété standard non seulement dans le milieu scolaire mais aussi médiatique et scientifique pour des considérations politiques et idéologiques d'inspiration fortement religieuse. Néanmoins, depuis quelques années, nous assistons grâce à la multiplicité des outils informatiques et à l'avènement de l'algorithmique (apprentissage profond) à des tentatives de documentation des variétés parlées qui ont été jusqu'à présent largement mises à l'écart.

Pour remédier à la non-accessibilité des données existantes et à la non-adaptabilité des corpus disponibles pour nos besoins, nous avons choisi de collecter un corpus du parler maghrébin, dans la perspective de mise à disposition. Animés par la volonté d'alignement sur un ensemble de bonnes pratiques déjà bien établies (Abouda & Baude, 2006 ; Baude et al., 2006 ; Gadet & Guerin, 2016 ; Wilkinson et al., 2016), notre méthodologie de collecte s'inspire de celle des ESLO (Enquêtes Sociolinguistiques à Orléans) (Baude & Dugua, 2016), adaptée aux spécificités du contexte maghrébin. Nous évoquerons la phase déterminante de conception du corpus qui comprend la définition des objectifs de la recherche et des types de données nécessaires (entretiens, conversations spontanées, etc.), ainsi que la méthodologie d'échantillonnage de locuteurs en termes de lieu géographique, de tranche d'âge, de genre et catégorie socio-professionnelle.

Ce corpus inclut des données variées, situées, protégées, transcrites, translittérées et segmentées en tours de parole. Il répond ainsi à une diversité de besoins et espère poser les bases de la constitution d'un corpus de référence et comparable entre les trois pays du petit Maghreb. Partagé entre les idées selon lesquelles la transcription serait une voie à éviter (Bird, 2020) et la réalité des pratiques dictées par la tradition et l'implantation forte d'outils d'analyses, nous proposons ici de reconsidérer une phase du développement de corpus oraux longtemps tenue pour subsidiaire par une étude exploratoire sur l'arabe parlé au petit Maghreb. Face aux interrogations impossibles à résoudre quant à la convention de transcription adéquate, l'interaction entre la transcription manuelle et les prédictions du système automatique dans le

processus itératif peut s’avérer bénéfique dans la prise de décision guidée par des contraintes pragmatiques.

Bibliographie

Abouda, L., & Baude, O. (2006). *CONSTITUER ET EXPLOITER UN GRAND CORPUS ORAL : CHOIX ET ENJEUX THEORIQUES. LE CAS DES ESLO*. https://halshs.archivesouvertes.fr/halshs-01162506

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* (No. arXiv:2006.11477). arXiv. https://doi.org/10.48550/arXiv.2006.11477

Baude, O., Blanche-Benveniste, C., Calas, M.-F., Cappeau, P., Cordereix, P., Goury, L., & Jacobson, M. (2006). *Corpus oraux, guide des bonnes pratiques 2006*. 209.

Baude, O., & Dugua, C. (2016). Les ESLO, du portrait sonore au paysage digital. *Corpus, 15*. https://doi.org/10.4000/corpus.2924

Bird, S. (2020). Sparse Transcription. *Computational Linguistics, 46*(4), 713- 744. https://doi.org/10.1162/coli_a_00387

Gadet, F., & Guerin, E. (2016). Construire un corpus pour des façons de parler non standard : « Multicultural Paris French ». *Corpus*, 15. https://doi.org/10.4000/corpus.3049

Peng, Z., Gerdes, K., & Guiller, K. (2022). Pull your treebank up by its own bootstraps. In

L. Becerra, B. Favre, C. Gardent, & Y. Parmentier (Éds.), *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)* (p. 139- 153). CNRS. https://hal.science/hal03846834

Plaquet, A., & Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. *24th INTERSPEECH Conference (INTERSPEECH 2023),* 3222- 3226. https://doi.org/10.21437/Interspeech.2023-205

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., & Auli, M. (2023). *Scaling Speech Technology to 1,000+ Languages* (No. arXiv:2305.13516). arXiv. https://doi.org/10.48550/arXiv.2305.13516

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A.,

Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Testing trade-offs between gender and number indexing and other cues to A and P arguments: A corpus based perspective

LÖFGREN Althea[1] and LEVSHINA Natalia[2]

1: MoDyCo, Université Paris Nanterre, CNRS
2: Radboud University

althea.margareta.lofgren@gmail.com

Mots-Clés : trade offs ; syntax ; agreement ; typology ; corpus linguistics

All languages have some formal and semantic cues that help the addressee understand who did what to whom", such as word order, case marking, agreement (indexing) and semantic and pragmatic properties of the referents. The trade-off relationship between word order and case marking has been well studied (Sapir 1921, Sinnemäki 2008, Koplenig et. al. 2017, Levshina 2021), but what role indexing plays is less known. This study aims at developing a new methodological perspective to investigate trade-offs between gender and number indexing and other cues to A and P arguments using data from corpora.

The body of research on trade-offs is characterized by two assumptions that are seldom explicitly tested. There is often an implicit assumption that language users try to communicate efficiently and that complexity in one area of language demands simplicity in another. Kemp & Regier 2012 find a trade-off between cognitive and communicative costs in lexical systems of kinship words or color terms and Koplenig et. al. (2017) demonstrate trade-offs between information conveyed by word order and word structure in Bible translations. Conversely, Levshina (2021) finds that in investigating the trade-off between cues for A and P arguments, there is little evidence for efficient language use and argues that the interaction between cues can best be explained by sociolinguistic factors. Similarly, Sinnemäki (2008) tests the trade-off theory on head/dependent marking and word order correlations and finds that, while all correlations were indeed negative, only two were statistically significant: the correlations between word order and dependent marking, and between word order and morphological marking (the sum of head and dependent marking).

The research goals are twofold. The first one is to develop a quantitative corpus-based measure of the contribution of indexing (agreement) to disambiguation of A and P. The second is to check if there are trade-offs (correlations) between this measure and other corpus-based variables which represent case marking, word order rigidity, semantic tightness and the position of the verb in a clause.

Nominal transitive clauses with gender and number indexing were obtained from the SUD treebanks, a collection of syntactically annotated corpora based on Universal Dependencies, for a sample of 30 cross-linguistically diverse languages (Gerdes et. al. 2018). The data was annotated to examine the influence of gender and number indexing on distinguishing A and P. A disambiguation index was developed to measure this impact, defining ambiguous A and P as having matching number and/or gender, and unambiguous A and P as differing in these aspects with A agreeing with the verb. Finally, we tested pairwise correlations between agreement and the other cues with the help of mixed models which include phylogenetic information and geographical distances between the sample languages.

The results reveal a significant trade-off between word order rigidity and number/gender indexing. Specifically, languages that rely heavily on rigid word order are less likely to employ number and gender disambiguation, while languages with more flexible word order are more likely to do so.

Bibliographie

[1] Gerdes, Kim, Guillaume, Bruno, Kahane, Sylvain, & Perrier, Guy. 2018 (Nov.). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In: Lynn, Teresa, & Schuster, Sebastian (eds), Universal Dependencies Workshop 2018.

[2] Kemp, Charles, & Regier, Terry. 2012. Kinship Categories Across Languages Reflect General Communicative Principles. Science (New York, N.Y.), 336 (05), 1049–54.

[3] Koplenig, Alexander, Meyer, Peter, Wolfer, Sascha, & Muller-Spitzer, Carolin. 2017. The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. PLoS ONE,12 (03), e0173614.

[4] Levshina, Natalia. 2021. Cross-Linguistic Trade-Offs and Causal Relationships Between Cues to Grammatical Subject and Object, and the Problem of Efficiency-Related Explanations. Frontiers in Psychology, 12. Sapir, E. 1921. Language: An Introduction to the Study of Speech. A Harvest book HB7. Harcourt, Brace.

[5] Sinnemäki, Kaius. 2008. Complexity trade-offs in core argument marking. Pages 67–88 of: Miestamo, M., Sinnemäki, K., & Karlsson, F. (eds), Language Complexity: Typology, contact, change. Studies in language companion series. John Benjamins.

Typologie de la perception des dynamiques urbaines dans des données multimodales : cas du projet d'Europacity

LESSIEUX Aurore

MoDyCo, Université Paris Nanterre, CNRS

aurore.lessieux@wanadoo.fr

Mots-Clés : opinion mining ; corpus oral ; transcription automatique ; perception ; urbanisme

1. Contexte de la recherche

Les dynamiques urbaines – transformations physiques, sociales, et économiques qui surviennent dans les zones urbaines (constructions, rénovations, évolutions, étalement urbain) – sont au cœur du projet ViTAL (Ville et Traitement Automatique du Langage). Il s'agit d'un projet pluridisciplinaire associant de manière originale la recherche et l'ingénierie à travers une collaboration innovante entre des chercheurs en TAL, des architectes et des urbanistes, en partenariat avec les Archives nationales, la BnF et l'INA. Ce projet a pour objectif de comprendre la manière dont les discours et les projets urbains abordent l'espace comme une dynamique. L'hypothèse posée est que pour comprendre la manière dont les sociétés se projettent dans l'espace, il est essentiel d'étudier des conceptions différenciées de la dynamique urbaine : repérer et analyser les éléments du discours reflétant la notion de dynamique urbaine et de sa perception par les acteurs de l'urbanisme dans des projets et initiatives d'aménagement.

Le projet VITAL s'appuie sur l'analyse des textes écrits par des acteurs de l'urbanisme (élus, techniciens, concepteurs) et se concentre sur la détection et l'analyse de la dynamique urbaine et de ses caractéristiques [1]. Ma recherche de doctorat s'inscrit dans les axes visés par le projet VITAL et se focalise sur l'étude de la perception des projets d'aménagement dans les données multimodales mises à disposition par l'INA. L'ambition de ma recherche est de développer un outil de classification supervisé [2] de la perception des changements urbains qui sera par la suite mis à disposition par le lab de l'INA dans leur catalogue d'outils de fouille et d'analyse automatisée spécifiquement adaptés aux corpus multimédias. Il est à noter que là où le projet VITAL accède uniquement à la perception d'un petit groupe d'acteurs de l'urbanisme, l'accès aux collections audiovisuelles de l'INA permet d'entendre la voix d'un plus grand nombre d'acteurs de l'urbanisme mais également d'acteurs difficile d'accès si ce n'est dans les productions journalistiques : les acteurs touchés de l'urbanisme (habitants, usagers, commerçants).

Dans cette communication, je souhaiterai présenter l'avancement de mes travaux, à savoir la construction du corpus multimodal ainsi que l'analyse et la modélisation de la perception de la dynamique urbaine.

2. Constitution de corpus

L'accès aux collections de l'INA a permis la construction de trois corpus de projet d'aménagement ayant fait l'objet de médiatisation : 1. le projet du Grand Paris initié par Nicolas Sarkozy en 2009, 2. le méga-projet urbain contesté Europacity, et 3. le projet Eurodisney qui permettra l'étude diachronique d'un projet d'aménagement. Ils sont constitués de contenus audiovisuels de nature variée – reportage, JT, micros-trottoirs, documentaires, débats politiques, interview – sélectionnés manuellement selon leur pertinence d'après les métadonnées renseignées par les documentalistes dans les fonds TV et Radio. Le corpus ainsi constitué comprend 1086 vidéos équivalent à environ 273 heures de contenus multimédias. Les technologies de transcription automatique ont permis un énorme gain de temps dans la création des corpus en étant source de données langagières. L'ensemble du corpus a été transcrit automatiquement, le premier avec l'ASR du LIUM [3] qui permet une segmentation du flux audio par locuteurs, et les deux autres avec WhisperX [4] qui permet une segmentation en tours de parole mais sans indication de locuteur. Malheureusement, la transcription automatique n'est pas infaillible et nécessite une correction manuelle pour pouvoir être exploitée dans cette recherche. La correction impliquait de reprendre les transcriptions pour assurer leur conformité avec l'audio : erreurs d'homophonie, suites de mots sans sens, bouts de transcription absente, fautes de syntaxe. Ces corrections ont été accompagnées de l'ajout d'éléments paralinguistiques, de l'alignement audio/transcription, de la segmentation en tours de parole et de leur attribution aux différents locuteurs regroupés par type d'acteur urbain.

Pour pouvoir analyser la perception des projets urbaines, il est nécessaire de connaitre le rôle du locuteur dans ces projets [5]. Le choix a été fait de distinguer trois groupes de locuteurs selon l'intensité de leur relation au projet et leurs rôles pour l'avancée du projet : 1. les acteurs extérieurs, qui apportent une supervision ou une influence indirecte sans intervenir activement dans la gestion quotidienne du projet (les associations, les mandataires, les journalistes), 2. les acteurs internes, qui jouent un rôle dans la réalisation concrète du projet (les décideurs, les techniciens, les concepteurs), et 3. les acteurs interne-externes, ceux que le projet affecte et qui vivent ou interagissent directement avec les résultats du projet (les habitants, les commerçants, les usagers).

3. Modélisation de la perception

La notion de perception est une des notions au centre des recherches dans le domaine de la psychologie cognitive. C'est par le biais des sens qu'un individu perçoit le monde [6]. Le traitement sensoriel des informations physiques, indépendamment de leur signification fait référence aux sensations [7]. Par la suite, ces sensations vont être interprétées subjectivement en se confrontant à des informations en mémoire : connaissances, croyances, expériences passées et empruntes émotionnelles associées [8]. La perception n'est pas une simple copie de la réalité. La perception des objets qui nous entourent s'effectue par la reconstruction cognitive de la réalité propre à chacun [9].

La perception est donc une expérience interne qui varie d'un individu à l'autre mais qui peut être captée lorsqu'un individu s'exprime. Les choix de lexique, des modes verbaux ou encore des constructions syntaxiques se font en fonction du regard que l'on porte sur l'objet du discours. Ces choix montrent la perception que le locuteur a de la réalité et la façon avec laquelle il veut construire sa représentation. Selon [10], [11], la perception comprend trois types d'informations : (1) l'information cognitive qui renvoie aux croyances du locuteur vis-à-vis de l'objet de son discours ; (2) l'information affective qui correspond aux réactions émotionnelles du locuteur suscitées par cet objet ; (3) l'information conative qui reflète ses intentions futures.

Nous considérons que l'information cognitive correspond à la notion du jugement, le discours qui vise à parler du monde :

(1) Je crois que les travaux de rénovation seront achevés dans les délais.

L'information affective renvoie vers les notions de sentiment et d'émotion :

(2) Je suis ravie que les travaux aient été achevés dans les délais.

Enfin, l'information conative peut être associée à la volonté, c'est-à-dire le discours qui cherche à changer quelque chose dans le monde :

(3) Les travaux de rénovation devront être achevés dans les délais.

4. Conclusion

La communication présentera les différentes étapes des traitements mises en place pour analyser la perception du projet d'aménagement Europacity. Il s'agit d'un ancien projet de mégacomplexe très contesté et médiatisé qui comptait rassembler un centre commercial, un parc de loisirs, des équipements culturels et des hôtels sur le territoire du « Triangle de Gonesse » mais qui fut abandonné en 2019 sous la pression des citoyens. Nous décrirons la transcription

et l’annotation de ce corpus. Nous présenterons quelques résultats issus de l’analyse quantitative et qualitative du corpus annoté.

Bibliographie

[1] Eshkol-Taravella, Iris & Jade Mekki & Olivier Ratouis & Alain Guez & Rémi Simon, « Repérer et caractériser les dynamiques urbaines : l'appui nouveau des humanités numériques », Humanités numériques [En ligne], 9 | 2024, mis en ligne le 01 juin 2024, consulté le 04 septembre 2024. DOI : https://doi.org/10.4000/11wmy

[2] Nguemegne, Emmanuelle Kelodjoue. Classification de transcriptions orales dans un contexte applicatif peu doté : application du TAL pour l'analyse de verbatim destinée à l'évaluation de l'acceptabilité d'une innovation. (2022) Traitement du texte et du document. Université Grenoble Alpes.

[3] Rouvier, Mickael & Dupuy, G. & Gay, P. & Khoury, E. & Merlin, T. & Meignier, S. (2013). An Open-source State-of-the-art Toolbox for Broadcast News Diarization. Interspeech, Lyon, France.

[4] Lau, H. & Michel, M. & LeDoux, J.E. & al.. (2022). The mnemonic basis of subjective experience. Nat Rev Psychol 1, 479-488.

[5] https://villedurabledotorg.wordpress.com/guide-de-gestion-de-projets-urbains/principes-strategiques-pour-la-gestion-de-projets-urbains/les-acteurs-du-projet-urbain-et-leurs-roles/, consulté le 30/05/2024.

[6] Flamein, Hélène, Etude de la perception d'une ville : Repérage automatique, analyse et visualisation. (2019) Thèse de doctorat en TAL, sous la direction d'Iris Eshkol-Taravella, Nanterre, Université Paris Nanterre.

[7] Lieury, A. & Léger, L. Chapitre 3. La perception du monde. (2020). Dans : , A. Lieury & L. Léger (Dir), Introduction à la psychologie cognitive, 61-84. Paris: Dunod.

[8] Gibson, James Jerome. (1966). The senses considered as perceptual systems.

[9] Bruner, J. S., & Goodman, C. C. Value and need as organizing factors in perception. (1947). The Journal of Abnormal and Social Psychology, 42(1), 33-44.

[10] Rosenberg, M. J. & Hovland, C. I. & McGuire, W. J. & Abelson, R. P., & Brehm, J. W. Attitude organization and change: An analysis of consistency among attitude components. (1960). Yales studies in attitude and communication. Yale Univer.

[11] Zanna, M. P., & Rempel, J. K. Attitudes: A new look at an old concept. (1988). In D. Bar-Tal & A. W. Kruglanski (Eds.), The social psychology of knowledge, 315-334. Cambridge University.

RAG pour l'exploration de corpus en GLAM

FACI Adam[1] and DE SACY Antoine Silvestre[2]

1 : Centre National de la Recherche Scientifique (CNRS)
2 : CNRS Lattice

adam.faci@live.fr, antoinedesacy@gmail.com

Mots-Clés : TAL ; LLM ; GLAM ; humanités numériques ; expérience utilisateur

Depuis de nombreuses années maintenant, les GLAMs [1] ont procédé à des politiques de numérisation massives de leurs collections, les rendant accessibles librement et sous la forme de données numériques aux chercheurs. La conséquence de cet accès massif aux données est de devoir imaginer de nouvelles manières de les parcourir, sous peine d'être submergé, que ce soit pour les chercheurs, ou les bibliothécaires et conservateurs qui souhaitent les valoriser.

En effet, ces fonds et leurs métadonnées constituent de grands volumes de connaissances que les techniques traditionnelles ne permettent pas forcément de traiter dans leur ensemble. À cela s'ajoute que certains fonds sont protégés et peuvent être exploités de façon limitée (collections sous-droits, notamment) [2].

Par ailleurs, le besoin de disposer de compétences pointues dans plusieurs domaines (informatique et analyse des données, infocom, expertise des corpus à l'étude, humanités numériques) nécessite de travailler au sein d'équipes pluridisciplinaires [2].

Ces outils sont, par exemple, les très populaires modèles de langue génératifs, fine-tunés pour le chat, tels que ChatGPT [5]. Cependant, ce sont des « boîtes noires », dont l'interprétation des sorties est limitée. C'est problématique à plus d'un titre : justifier des résultats de recherche et les contextualiser par des références à l'état de l'art est crucial.

Enfin, les corpus sont servis à différents publics [6]. Cette multiplicité des destinataires demande d'adapter les corpus selon l'objectif, les qualifications ou encore les droits d'accès.

Le Retrieval Augmented Generation (RAG) [] est une technique utilisée en couplage avec les modèles génératifs [5]. La génération est contextualisée par un corpus de documents sur lequel le modèle génératif va s'appuyer pour fournir ses réponses : des extraits sont sélectionnés par un autre modèle et fournis au modèle génératif dans le prompt. En sortie, ces extraits sont donnés comme contexte de la génération. Ainsi, selon le contexte, c'est-à-dire selon le corpus utilisé et les extraits sélectionnés, la réponse est différente. On peut alors varier le domaine (médecine, littérature, droit) ou la temporalité (corpus ancien ou récent).

Un autre avantage est que, contrairement à un entraînement de modèle qui nécessite des ressources et un travail d'organisation des connaissances, le RAG est plus aisé et accessible.

De plus, les données d'entraînement ne sont pas disponibles pour la majorité des grands modèles de langue : RAG assure un certain niveau de transparence par la mise en évidence de passages déterminants pour la génération.

Pour finir, le RAG permet d'extraire localement des informations en usant de petits modèles qui ne nécessitent pas d'authentification ou de transfert de données : il est utilisé le temps de la tâche, ses connaissances du corpus ne mettent pas à jour le modèle, et il peut donc être exécuté sur des données personnelles ou sensibles.

Dans la perspective des GLAMs, et dans le contexte d'exploration de corpus avec un moteur de recherche, par exemple pour la constitution d'un état de l'art et son exploration, plusieurs stratégies utilisant le RAG peuvent être mises en place. Un obstacle reste, à partir d'un même ensemble de documents, d'établir des sous-corpus permettant de disposer de différents contextes à fournir au RAG.

Notre proposition est d'associer des techniques de détection de communautés (autour d'auteurs influents et autour de thématiques) [Ref clusterisation][6] et de classification de ces communautés [Ref] pour les décrire en amont. Le couplage de cette détection de communautés à l'utilisation du RAG permet ainsi de présélectionner et donner à voir un premier filtrage du corpus à l'utilisateur sur lequel le RAG peut fonctionner avec plus de pertinence. Lors de sa recherche documentaire, l'utilisateur peut ainsi voir se dessiner sous ses yeux des communautés et des réseaux entre les résultats de sa recherche, réseaux qu'il pourra extraire et sur lesquels il pourra utiliser le RAG pour questionner ces communautés, sans jamais perdre la main sur cette constitution.

Nous proposons également d'autres chaînes de traitements permettant, de manière interactive, de composer différentes communautés dans un corpus, de les caractériser à partir de critères spécifiques : thème caractéristique du cluster, richesse lexicale, niveau de technicité, etc.

Nous proposons ainsi d'utiliser le RAG de différentes façons :

(a) en contextualisant la réponse ;

(b) en fournissant à l'utilisateur les données déterminantes pour la génération dans le contexte ;

(c) en organisant les réponses sous la forme de séries de contextualisation. Dans ce dernier cas, une réponse correspondant à un contexte, cela permet à l'utilisateur d'analyser les spécificités des différents clusters.

D'autres stratégies sont également à explorer, notamment en permettant de varier le choix du modèle de sélection et ses paramètres, permettant de retrouver différentes communautés.

Aussi, penser l'interactivité d'un tel processus passe par la question des éléments à présenter à un utilisateur, l'ordre de leur présentation et les différentes vues sur ces présentations.

Nous proposons alors d'expérimenter sur ces différentes vues et établir des cas d'usages en fonction de profils utilisateurs prédéfinis.

Bibliographie

[1] E. Bermes, «Le numérique en bibliothèque: naissance d'un patrimoine: l'exemple de la Bibliothèque nationale de France (1997-2019),» Thèse de l'École nationale des chartes – PSL, 2020.

[2] G. K. e. C. Scopsi, «La recherche ouverte et les données en Lettres, Sciences humaines et sociales (LSHS),» chez Les nouveaux paradigmes de l'archive, Publications des Archives nationales, 2024, pp. 127-143.

[3] P. e. a. Lewis, «Retrieval-augmented generation for knowledge-intensive nlp tasks.,» chez Advances in Neural Information Processing Systems 33, 2020.

[4] J.-F. Bert, «Pratiques d'archives: Problèmes actuels sur les usages du matériau documentaire.,» chez La diffusion numérique des données en SHS - Guide de bonnes pratiques éthiques et juridiques, Digitales, 2018, pp. 31-38.

[5] M. K. e. D. Dwivedi, «Emotional AI: Neuroethics and Socially aligned networks,» chez Emotional AI and Human-AI Interactions in Social Networking, Muskan Garg et Deepika Koundal, 2023, pp. 101-130.

[6] G. Maingot, «Enquête Observatoire des publics de la BnF,» Bibliothèque Nationale de France, 2016.

[7] M. C. e. A. Laborderie, «La BnF et les services à la recherche à l'heure des humanités numériques,» Arabesques, vol. 105, pp. 8-9, 2022.

[8] S. Fortunato, «Community detection in graphs,» Physics Reports, vol. 486, pp. 75-174, 2010.

Session Poster

Méthodologie de collecte et de compilation d'un corpus à partir du site YouTube : le cas des vidéos de présentation d'expositions de mode britanniques et américaines

GANET Agnès

Centre de Linguistique Inter-langues, de Lexicologie, de Linguistique Anglaise et de Corpus, Université Paris Cité
Langues, Enseignement et Anglais De Spécialité, Ecole Normale Supérieure Paris-Saclay

agnes.ganet@ens-paris-saclay.fr

Mots-Clés : Linguistique de corpus ; anglais de spécialité ; YouTube ; numérique ; mode ; expositions ; musées

Les expositions de mode, nées au lendemain de la Seconde Guerre Mondiale, ont connu depuis 2008 une croissance exponentielle [1] . Elles occupent désormais une place importante dans le paysage muséal [2] [3] et sont très populaires [4] [5] [6].

Dans le cadre du développement des nouvelles technologies, des discours qui naissaient et circulaient autrefois traditionnellement et exclusivement par écrit et au format papier se voient désormais migrer vers le Web [7]. C'est par exemple précisément le cas des expositions de mode, qui peuvent désormais être présentées dans le cadre de vidéos postées par les musées sur leur chaîne YouTube. Si, auparavant, les expositions ne pouvaient être présentées qu'à l'écrit (que ce soit dans des journaux, des magazines, des communiqués de presse ou encore des catalogues), elles ont désormais la possibilité d'être présentées sur la plateforme de partage de vidéos YouTube. Ces vidéos apparaissent alors comme des opportunités nouvelles [8] [9], qu'il convient d'étudier linguistiquement au vu de l'influence de ce réseau social.

Ainsi, les pratiques muséales ont progressivement intégré les réseaux sociaux et les nouvelles technologies [1], comme de nombreux autres domaines spécialisés [10] [11]. YouTube est de plus en plus perçu par les musées comme un outil important pour communiquer avec le public [12]. Deuxième site le plus visité du monde et premier site le plus visité dans les catégories « Arts et divertissement » d'après le site « Top websites ranking » [13], YouTube est très populaire [14], et s'impose désormais comme une plateforme privilégiée pour partager des contenus numériques nouveaux et les diffuser à un large public.

De fait, les vidéos de musées présentant des expositions de mode représentent un « candidat-genre spécialisé » [15], puisqu'aucune étude en anglais de spécialité ne s'est attachée, à notre connaissance, à en décrire les figements linguistiques et rhétoriques, ni à établir le degré de spécialisation de ce type de discours. Ce candidat genre numérique récent est natif du Web

(« émergent », selon la terminologie de Herring 2013 [16]), et oral, et présente à ce titre des spécificités discursives. YouTube apparaît alors comme un véritable vecteur de production langagière et est, à l'image du Web, un « répertoire de genres » [7].

Par conséquent, YouTube est devenu une ressource précieuse et privilégiée pour constituer des corpus oraux et numériques : cet outil technologique représente désormais une source importante de données langagières et de corpus potentiels. Les vidéos de présentation d'expositions de mode postées sur YouTube représentent un nouveau type de discours multimodal, généré dans cet espace numérique, et est propre à cet espace.

Ainsi, étudier le langage utilisé dans ces vidéos invite à s'interroger sur la méthodologie de la constitution de corpus numériques et oraux. Ces derniers sont encore peu explorés en anglais de spécialité [17], et il y a, en règle générale, « peu de corpus oraux » [18]. Afin de mener ce travail, nous avons compilé notre propre corpus, qui est constitué de transcriptions de vidéos YouTube en anglais. En effet, à notre connaissance, aucun corpus de transcriptions de vidéos d'expositions de mode n'existe, dans quelque langue que ce soit, comme cela est souvent le cas pour les corpus spécialisés [19]. Nous avons donc dû choisir avec précaution les éléments qui le constituent et établir nos propres critères de sélection [19]. Nous nous sommes appuyée sur la liste d'expositions de mode du site 'Exhibiting Fashion' [20], puis avons poursuivi nos recherches pour trouver d'autres vidéos non répertoriées sur ce site. Les vidéos doivent avoir été postées entre 2008 et 2023 par des musées britanniques et américains. Il s'agit d'un objet d'étude original, tant parce que le sujet est nouveau et non exploré en anglais de spécialité, mais aussi parce que son moyen de diffusion, YouTube, est lui aussi récent [21].

Néanmoins, de nombreuses difficultés apparaissent lors de la compilation de corpus oraux [22] et la constitution de notre corpus n'a pas été sans obstacles. L'utilisation de l'outil « transcription » de YouTube a aidé à compiler ce corpus. Toutefois, la qualité des transcriptions n'est pas optimale, et a requis des vérifications systématiques et minutieuses. Nous avons donc établi une typologie d'erreurs de transcription. Par exemple, la transcription de noms propres (entités nommées), de mots étrangers, de termes spécialisés et la confusion entre les -s du pluriel et les -s du génitif sont tant d'exemples d'erreurs fréquentes dans les transcriptions YouTube. Cet outil est donc très utile en termes de compilation de corpus, mais n'est pas sans poser de difficultés, et on ne saurait suffisamment insister sur « l'ampleur de la tâche de transcription » [18]. Une autre question cruciale est celle des méta-données, telles que la date de publication de la vidéo, le musée qui en est à l'origine, ou encore le type de vidéo. Ces méta-données permettent de documenter le corpus et d'ensuite interroger différentes parties du corpus pour pouvoir effectuer des statistiques.

L'outil numérique qu'est YouTube est donc utilisé pour rassembler des données avec pour but, à terme, de caractériser le langage oral spécialisé de la mode en contexte muséal avec une approche corpus-driven [19]. À ce corpus oral et numérique s'ajoute un corpus de référence écrit [23], préalablement constitué, qui rassemble différents genres discursifs spécialisés dans le domaine des expositions de mode (communiqués de presse, articles de presse et de magazines, cartels d'expositions et catalogues). Celui-ci permet d'ores et déjà de disposer de ressources terminologiques et phraséologiques afin de dégager les spécificités de notre corpus numérique. En outre, nous avons ainsi pu constater des différences méthodologiques dans la manière d'aborder ces corpus différents, car la complexité du traitement des données orales entraîne la nécessité d'adopter une méthodologie spécifique.

Nous présenterons ainsi la méthodologie de collecte de corpus oral que nous avons adoptée, par opposition à la collecte du corpus de référence écrit. Il s'agira aussi d'expliciter les choix de méta-données effectués. Nous proposerons ensuite une typologie d'erreurs de transcription, telles que des termes ou des noms propres erronés, ainsi que les stratégies développées pour permettre de corriger ces erreurs. Nous formulerons ensuite des hypothèses quant aux visées discursives promotionnelles et informatives du genre et de la manière dont cela se traduit linguistiquement

Bibliographie

[1] Mida, I. (2015). The Enchanting Spectacle of Fashion in the Museum. Catwalk: The Journal of Fashion, Beauty, and Style, 4(2), 47-70.

[2] Palmer, A. (2008). Reviewing fashion exhibitions. Fashion Theory, 12(1), 121-126.

[3] Melchior, M. R. (2011, September). Fashion museology: Identifying and contesting fashion in museums. In Full paper draft, presented at Fashion. Exploring critical issues conference, Mansfield College, Oxford.

[4] 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.

[5] Petrov, J. (2019). Fashion, history, museums: Inventing the display of dress. Bloomsbury Academic.

[6] 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.

[7] Santini, M. (2006). Interpreting genre evolution on the Web. In Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources.

[8] Ansori, M., & Taopan, L. L. (2019). A multimodal discourse of promotional video wonderful Indonesia. English and Literature Journal, 6(1), 1-18.

[9] Xia, S. (2023). Explaining science to the non-specialist online audience: A multimodal genre analysis of TED talk videos. English for Specific Purposes, 70, 70-85.

[10] Hamilton, C. E. (2022). L'impact des corpus numériques sur la transmission et l'évaluation des connaissances en didactique des langues. Numérique et didactique des langues et cultures: Nouvelles pratiques et compétences en développement, 37-54.

[11] Birch-Becaas, S., Kloppmann-Lambert, C., Carter-Thomas, S., Dressen-Hammouda, D., Rowley-Jolivet, E., & Zerrouki, N. (2023). Research dissemination in digital media: An online survey of French researchers' practices. ASp. la revue du GERAS, (84), 113-136.

[12] 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.

[13] 'Top Websites Ranking' (2023). Similarweb, consulté le 8/12/2023 <https://www.similarweb.com/top-websites/> et <https://www.similarweb.com/top-

websites/> [14] Godwin-Jones, R. (2007). Digital video update: YouTube, flash, high-definition.

[15] Kloppmann-Lambert, C. (2023). Style spécialisé et styles personnels dans le genre des billets de blog professionnel d'architecte. Études de stylistique anglaise, (17).

[16] Herring, S. C. (2013). Discourse in Web 2.0: Familiar, reconfigured, and emergent. Discourse, 2(0), 1-26.

[17] Gautier, L., & Hohota, V. (2014). Construire et exploiter un corpus oral de situations de dégustation: l'exemple d'Oenolex Bourgogne. Studia Universitatis Babes-Bloyai, Philologia, 59(4), 157-173.

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

[19] 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).

[20] 'Exhibiting Fashion', Centre for Fashion Curation (2024), University Arts London.
https://fashionexhibitionmaking.arts.ac.uk/archive/

[21] Mady, M. A., & Baadel, S. (2020). Technology-Enabled Learning (TEL): YouTube as a ubiquitous learning aid. Journal of Information & Knowledge Management, 19(01), 2040007.

[22] Adolphs, S., & Knight, D. (2010). Building a spoken corpus. The Routledge handbook of corpus linguistics, 38-52.

[23] Ganet, A. (2023). L'itinéraire textuel des expositions de mode, des premiers communiqués de presse aux catalogues d'exposition: étude du tétraptyque information-didactique-esthétisme- promotion (mémoire de Master 2, sous la direction de Natalie Kübler)

Presenting LongFoRMer:
A package to organize and analyze long-form recordings

TEY Kai Jia[1], PEUREY Loann[1], DAS Shuvayanti[1], GAUTHERON Lucas[1,2], HAVARD William[1,2], SCAFF Camila[1,4], CRISTIA Alejandrina[1]

1: Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'études cognitives, École normale supérieure, PSL Research University, CNRS
2: University of Wuppertal
3: University of Orléans
4:University of Zurich

kaijiatey@gmail.com

Mots-Clés : Long form recordings ; Automatic Annotation ; Data Management and Standardisation

Long-form recordings (LFR) collected via child-worn devices are becoming increasingly common in children's input and production studies (see: Karadayi et al., 2018; Scaff et al., 2024). The wearable devices capture what children hear and say over the course of an entire day, resulting in large data files. However, managing these data poses several technical and usability challenges, especially because of the sensitivity and sheer volume of the data. Many researchers struggle to manage these massive datasets, such as having multiple copies of the same large audio files or having divergent spreadsheets describing extracts from the audio files, often with varying naming conventions or annotation schemes.

We have developed LongFoRMer (Long-form Recording Manager, formerly ChildProject), a package that allows researchers using LFR to organize their files in a standardized way, ensuring consistent data management and facilitating interoperability across systems. This package also provides procedures to import annotations from a wide range of existing formats (LENA's .its, ACLEW annotation structure in ELAN, Praat) into standardized .csv files. It includes clever solutions for the above-mentioned problems, such as annotations covering only sections of the audio and/or subsets of the participants. The stored data files can be automatically converted to a wide range of different formats. These formats will retain their link to the original data file while providing a converted version that can be used for other analytical purposes.

Besides, LongFoRMer can extract descriptive statistics on language input and production from the annotations, such as the number of vocalizations, and average vocalization length by different speakers. These metrics are important for measuring the relationship between

children's speech production and adult speech (Bergelson et al., 2023). Through LongFoRMer's standardized organization, researchers can also benefit from straightforward instructions to apply open-source and free automated algorithms to return adult word counts and child vocalization counts. The package also includes procedures to evaluate the reliability of automated annotations against their human equivalents. After working with several labs in their exploration of our package, we have developed improved tutorials and troubleshooting sessions.

The package includes pipelines to facilitate connection with citizen science platforms (Zooniverse), which can greatly assist in gathering annotations from annotators worldwide for human manual coding of the data. LongFoRMer selects, prepares and uploads the data to the platform while keeping a record of the uploaded data and its linked metadata. LongFoRMer also includes a set of features that can select portions of audio of interest using specific selection criteria, such as periodic selection, high vocal activity or high amount of speaker interaction. These selections generated files are then ready for annotation using specialized software like ELAN.

Finally, the package relies on open-source tools that facilitate other aspects of work with LFR, namely Datalad, which allows lighter versions of the data (by not including the recordings), and GIN, which helps keep track of dataset versions and control sharing and collaboration. Putting everything together, the package offers an economical, standardized, not language-specific alternative to proprietary software.

Bibliographie

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., R. Hamrick, L., Marklund, E., Kalashnikova, M., Guez, A., & Casillas, M. (2023). Everyday language input and production in 1,001 children from six continents. Proceedings of the National Academy of Sciences, 120(52), e2300671120.

Karadayi, J., Scaff, C., Stieglitz, J., & Cristia, A. (2018). Diarization in Maximally Ecological Recordings: Data from Tsimane Children. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages, SLTU 2018, 30–35. https://doi.org/10.21437/SLTU.2018-7

Scaff, C., Casillas, M., Stieglitz, J., & Cristia, A. (2024). Characterization of children's verbal input in a forager-farmer population using long-form audio recordings and diverse input definitions. Infancy, 29(2), 196–215.

Bilingual Corpus Building with OpenAI's Whisper for Persian-English

NAMDARZADEH Behnoosh

CLILLAC-ARP Université Paris Cité

behnoosh.namdarzadeh@etu.u-paris.fr

Mots-Clés : Corpus linguistics ; OpenAI's Whisper ; Under ; resourced language ; Universal Dependencies ; Persian ; English bilingual dataset

This paper discusses how an audio Large Language Model (LLM) like Whisper (Radford et al., 2023) can be used to build spoken corpora for Persian. The aim is to present an overview of linguistic corpus building in a context of technological changes, replicating what has been done for Hebrew (Marmor et al., 2023) and Mandarin (Sun et al., 2024). I report first-hand experience on collecting spoken data for an under-resourced language for NLP tasks: Persian (Freihat & Abbas, 2021; Taghizadeh & Faili, 2016). Languages are classified as under-resourced when they lack the quantity of data necessary for training statistical and machine learning models (Liu et al., 2022). “Transcription bottlenecks” for under-resourced and endangered languages lead me to leverage the End-to-End Automatic Speech Recognition (ASR) system for my corpus collection of Persian (Shi et al., 2021; Zahrer et al., 2020). My PhD aims at creating a UD model to train spoken Persian data to detect dislocations. Next, I use Whisper to translate spoken Persian data into English, in order to create a bilingual corpus to be used as training data for Machine Translation (MT) systems. This may appear circular, however, I have to emphasize on the fact that Whisper is primarily being used for speech recognition and not translation. Its role is to transcribe spoken Persian data into text, which can then be aligned with pre-existing English text. Furthermore, human verification and corrections will be incorporated to ensure that the parallel corpus maintains high-quality standards. For some language pairs, including Persian-English, due to the scarcity of parallel data, this could be mentioned as a bootstrapping approach, where the initial data, though imperfect, serves as a starting point for further improvement of the MT system.

We will first discuss the scarcity of NLP resources for Persian, especially spoken corpora in informal context. Persian is diglossic and resources are even rarer for informal spoken data (Kabiri et al., 2022). This can pose challenges in various fields like Neural MT (NMT) and ASR. Notably, while there is a plethora of corpora for tasks involving both written and spoken data, the prevalence of written data in training sets tends to overshadow spoken data, impacting the treatment of certain linguistic structures like dislocation. Despite the existence of corpora

catering to Persian, those focusing on spoken registers primarily derive from scripted sources such as films or series, lacking authenticity in real conversational contexts. However, efforts like the Persian Speech Corpus (Persian Speech Corpus, 2016) and ShEMO (Mohamad Nezami et al., 2019) database offer valuable resources for non-commercial use, aiming to address these limitations by providing aligned speech data and emotion-labelled utterances for research purposes, respectively.

Inspired by interdisciplinary approaches to artificial intelligence, I am eager to explore innovative methods that can advance corpus building in the field of Linguistics. By harnessing the capabilities of Whisper, I aim to delve into the creation of corpus for an under-resourced language, Persian, based on the collection of YouTube videos of dyads spontaneous conversations. Given the time-consuming nature of transcribing spoken data, especially YouTube videos where speakers speak spontaneously, I have used OpenAI's Whisper to build a corpus for Persian.

As part of my PhD project, I am currently focusing on leveraging a customised C++ implementation of Whisper, based on Radford (Radford et al., 2023). The process involves feeding an audio file into the system, which then identifies the language and proceeds with transcription, starting from a small model and progressing to the largest model available (large-v1 for Persian). The resulting transcriptions, in various formats such as .txt, .srt, are converted into a TextGrid format using the pysrt Python library (https://pypi.org/project/pysrt/) for easier readability and correction in Praat (Boersma & Weenink, 2022). The correction does not take as much time as transcribing from scratch. However, to give an idea of the time spent on correcting timestamps and transcription, it would take about two minutes per minute of audio. It is worth noting that this is not consistent across all audio files, as some require less correction. The final aim is the creation of a more accurate UD model (https://universaldependencies.org/) for spoken Persian, particularly in detecting a challenging Dependency Relation, dislocated. Moreover, Whisper can produce English translations of the transcriptions, allowing for the creation of a parallel Persian-English corpus for fine-tuning pre-trained models to leverage translation of spoken data.

Using the C++ implementation of Whisper offers two key advantages. Firstly, it generates a dictionary of all Whisper subtokens during audio transcription, helping in the analysis of potential errors in a 24-hour training dataset for Persian by understanding the subtokenisation process. Secondly, the Whisper.cpp implementation by Gerganov (https://github.com/ggerganov/whisper.cpp) allows for fast processing of Whisper parameters and visualization of subtokens and their associated probabilities. This allows users to access

probability scores for each subtoken, enabling calibration curves to report on the accuracy of the Whisper models. Last, this Whisper implementation can be used for language detection. Preliminary findings on the transcription of Persian in the large model of Whisper, which outperforms other models (for Persian), suggest that there are still some transcription issues. I observed confusions between the consonants and vowels when places of articulation are the same or very close, such as /t/ and /d/. It seems that Whisper models are trained on read speech, resulting in a tendency to generate formal variants of words. Consequently, Whisper models tend to favour the written variant of the token. When delving into the subtoken dictionary for Persian, it seems that the decoder is used to post-process the data, as evidenced by the fact that the internal representations of the Whisper models in the dictionary are only the isolated forms of graphemes. When delving into the subtoken dictionary for Persian, it becomes apparent that the decoder serves the purpose of post-processing the data. This is indicated by the presence of only isolated forms of graphemes in the internal representations of the Whisper models within the dictionary.

The aim of creating this spontaneous corpus for my PhD project is twofold. First, it seeks to enrich the Persian corpus for NLP tasks (such as UD annotation), and to ensure that linguistic labels relevant to spoken data, such as dislocated, are accurately annotated. Second, it aims to potentially improve neural machine translation and speech translation through fine-tuning.

Bibliographie

Boersma, P., & Weenink, D. (2022). Praat: Doing phonetics by computer (Version 6.2.06) [Computer software]. https://www.praat.org

Freihat, A. A., & Abbas, M. (2021). Proceedings of The Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021}. Proceedings of the 4th International Conference on Natural Language and Speech Processing: Workshop on NLP Solutions for Under Resourced Languages.

Kabiri, R., Karimi, S., & Surdeanu, M. (2022). Informal Persian Universal Dependency Treebank. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 7096–7105). European Language Resources Association. https://aclanthology.org/2022.lrec-1.768

Liu, Z., Richardson, C., Hatcher, R., & Prud'hommeaux, E. (2022). Not always about you: Prioritizing community needs when developing endangered language technology. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 3933– 3944). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.272

Marmor, Y., Misgav, K., & Lifshitz, Y. (2023). ivrit.ai: A Comprehensive Dataset of Hebrew Speech for AI Research and Development (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2307.08720

Mohamad Nezami, O., Jamshid Lou, P., & Karami, M. (2019). ShEMO: A large-scale validated database for Persian speech emotion detection. Language Resources and Evaluation, 53(1), 1–16. https://doi.org/10.1007/s10579-018-9427-x

Persian Speech Corpus. (2016). https://fa.persianspeechcorpus.com/

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, 202, 28492--28518. https://doi.org/10.48550/ARXIV.2212.04356

Shi, J., Amith, J. D., Castillo García, R., Guadalupe Sierra, E., Duh, K., & Watanabe, S. (2021). Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yolóxochitl Mixtec. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 1134–1145. https://doi.org/10.18653/v1/2021.eacl-main.96

Sun, J., Wu, Y., Audibert, N., & Adda-Decker, M. (2024). Création d'un corpus parallèle de styles de parole en mandarin via l'auto-transcription et l'alignement forcé. In Actes des 35èmes Journées d'Études sur la Parole, 291–300.

Taghizadeh, N., & Faili, H. (2016). Automatic Wordnet Development for Low-Resource Languages using Cross-Lingual WSD. Journal of Artificial Intelligence Research, 56, 61–87. https://doi.org/10.1613/jair.4968

Zahrer, A., Zgank, A., & Schuppler, B. (2020). Towards Building an Automatic Transcription System for Language Documentation: Experiences from Muyu. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 2893–2900). European Language Resources Association. https://aclanthology.org/2020.lrec-1.353

Reddit et les hommes : étude linguistique des formations politiques antiféministes

SERISIER Marie[1,2]

1: Quantitative and Digital Humanities Lab
2: Centre de Linguistique Inter-langues, de Lexicologie, de Linguistique Anglaise et de Corpus, Université Paris Cité

marie.serisier@ulb.be

Mots-Clés : TAL ; manosphère ; CMC ; linguistique de corpus

La pensée masculiniste contemporaine, définie comme “un mouvement social conservateur ou réactionnaire qui prétend que les hommes souffrent d’une crise identitaire parce que les femmes en général, et les féministes en particulier, dominent la société et ses institutions” (Dupuis-Déri, 2009 : 97) trouve ses origines dans les années 1970, époque de luttes pour l'égalité entre les genres aux Etats-Unis. Cette époque est marquée par une remise en question globale des rôles typiquement masculins et féminins de répartitions des tâches dans les sphères domestique et politique. Cet élan commun, porté par les femmes et les hommes s'est érodé dans les années 1980, avec l'arrivée de Ronald Reagan à la tête du pays, marquant un recul des droits sociaux et des subventions allouées aux centres d'IVG aux Etats-Unis.

La démocratisation d'Internet et la popularisation des algorithmes de recommandations personnalisés creuse encore l'écart avec le projet féministe initial et font d’Internet le premier creuset de cette pensée réactionnaire (Ging, 2017). Le relatif anonymat qu’offre Internet permet une désindividualisation et une polarisation des opinions (Lee, 2007) ouvrant la voie à des parcours de radicalisation et d’action violentes (Munn, 2019). Cet espace partagé devient alors le terrain d’une croisade pour les défenseurs des droits des hommes, dans une perspective de défense identitaire (Dupuis-Déri, 2015). C’est ainsi qu’en 2014, des journalistes et développeuse de jeux vidéo ont été les victimes d’un déchaînement de haine en ligne de la part des internautes masculins de 4Chan, dans ce que l’on a ensuite appelé le #Gamergate (Massanari, 2017).

Internet devient alors un espace privilégié de reconfiguration politique et la frontière entre les mondes sensible et virtuel se fait de plus en plus mince. Selon le recensement SimilarWeb[3], la plateforme Reddit est le treizième site le plus visité au monde. Il est compartimenté en

[3] SimilarWeb (2023) https://www.similarweb.com/fr/

communautés d'intérêt (subreddits) sur lesquels il est possible d'échanger des messages textuels et visuels (mèmes, GIF, émoticônes).

Nous nous proposons d'étudier quatre subreddits présents sur la plateforme : r/MensLib, r/Mensrights, r/thePurplePillDebate, r/TheRedPill. Notre terrain d'étude principal réside dans les trois communautés proposant les stigmates d'une appartenance à la mouvance globale antiféministe en ligne (la manosphère), r/MensRights, r/ThePurplePillDebate, r/TheRedPill. Nous comparons les résultats obtenus avec ceux de r/MensLib, une communauté proposant une perspective progressiste.

Le corpus étudié rassemble l'intégralité des soumissions principales et des commentaires des subreddits susmentionnés, c'est-à-dire un corpus composé de 518 894 soumissions principales et 15 194 548 commentaires. Les bornes temporelles du corpus s'étendent entre la création de chaque subreddit et la fin de l'année 2022. Chacune de ces communautés comporte au moins 90 000 membres.

Nous cherchons alors à interroger ce corpus de deux façons différentes :

- RQ1 : Comment caractériser le fonctionnement (intra et extra-communautaire) de ces communautés qui proposent un discours sur le genre militant et par conséquent un discours politique ?
- RQ2 : Dans une perspective de discours politique, quelles articulations entre l'individu et le collectif ?

Pour ce faire, nous utilisons la grille d'analyse de Freelon (2010) qui propose de caractériser les communautés politiques sur le web en trois catégories : libéral individualiste, communautaire et délibératif. Le modèle libéral individualiste est basé sur la mise en avant d'intérêts privés dans une perspective de non-coopération. Le discours communautaire cherche à creuser l'écart entre le groupe concerné et l'extérieur en favorisant les opinions du groupe d'intérêt. Enfin le modèle délibératif se veut plus transversal, ayant une capacité à ne pas dévier d'un sujet donné et en posant des questions non-rhétoriques. Ce cadre d'analyse a récemment été utilisé par des chercheur·ses sur la manosphère (Krendel, 2020 ; Wright et al, 2020). Les variables présentes dans cette grille d'analyse seront évaluées à l'échelle du corpus et de façon diachronique selon un protocole d'étude mêlant des méthodes quantitatives et qualitatives réplicables dans d'autres contextes de recherche. Cela nous permet de cerner plus précisément la façon dont le discours de ces communautés est une re-politisation de la question de la masculinité à l'ère du parler-numérique.

Dans un contexte de méthode hybride, l'aspect de linguistique de corpus est mis en avant par l'étude approfondie des mots-saillants de chaque communauté et des contextes d'énonciations (Partington, 2013). Nous nous pencherons notamment sur la façon dont les membres de ces communautés se racontent eux-mêmes, les anecdotes proposées ainsi que les stratégies argumentatives mises en place. Cette méthode de fouille de texte est associée à des méthodes issues du traitement automatique des langues nous permettant de dégager de modéliser les échanges entre les communautés du corpus (analyse de réseaux). Le travail d'analyse est effectué grâce à deux outils : langage de programmation R ainsi que le logiciel de traitement de corpus Sketch Engine.

Parmi les résultats préliminaires, nous avons observé une fragmentation idéologique importante des communautés r/TheRedPill et r/PurplePillDebate par rapport au reste du corpus, avec une langue commune plus éloignée de la langue générale, qui suggère probablement une appartenance à une approche communautaire. L'étude des mots-saillants nous suggère que toutes les communautés antiféministes du corpus proposent des champs sémantiques appartenant au rapport au corps, au développement personnel ou à l'affrontement. Le corps est l'objet qui focalise les efforts des membres de r/TheRedPill, dans une quête de dépassement de soi. La guerre est présente chez les membres r/MensRights dans leur vision des figures antagonistes notamment à travers des associations de termes comme : « feminists + attack » ou « sjw[4] + invade ».

Il est difficile de conclure à une fragmentation totale de r/TheRedPill et r/PurplePillDebate du reste des communautés, 358 utilisateurs ont échangé sur les quatre subreddits. De plus, il existe des liens entre r/MensLib et les autres communautés : 1656 utilisateurs communs avec r/PurplePillDebate, 5505 utilisateurs communs avec r/MensRights et 445 avec r/TheRedPill. r/MensLib est donc une communauté moins isolée de la manosphère qu'anticipé. Il s'agit désormais d'investiguer plus en détail la nature de ces liens : amicaux ou antagonistes.

[4] Sjw : acronyme pour « social justice warrior », militant·e progressiste.

Bibliographie

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. In InterJournal: Vol. Complex Systems (p. 1695). https://igraph.org

Dupuis-Déri, F. (2009). Le « masculinisme » : Une histoire politique du mot (en anglais et en français). Recherches féministes, 22(2), 97-123. https://doi.org/10.7202/039213ar

Dupuis-Déri, F. (2012). Le discours de la « crise de la masculinité » comme refus de l'égalité entre les sexes : Histoire d'une rhétorique antiféministe. Cahiers du Genre, 52(1), 119- 143. https://doi.org/10.3917/cdge.052.0119

Dupuis-Déri, F. (2015). Les antiféminismes : Analyse d'un discours réactionnaire. https://www.academia.edu/15180915/Les_antif%C3%A9minismes_analyse_dun_discours_r %C3%A9actionnaire

Freelon, D. G. (2010). Analyzing online political discussion using three models of democratic communication. New Media & Society, 12(7), 1172- 1190. https://doi.org/10.1177/1461444809357927

Ging, D. (2019). Alphas, Betas, and Incels : Theorizing the Masculinities of the Manosphere. Men and Masculinities, 22(4), 638- 657. https://doi.org/10.1177/1097184X17706401

Kilgarriff A, Rychlý P, Smrž P, Tugwell D. The Sketch Engine. Proceedings of the 11th EURALEX International Congress: 105-116, 2004.

Krendel, A. (2020). The men and women, guys and girls of the 'manosphere': A corpus-assisted discourse approach [Publisher: SAGE Publications Ltd]. Discourse & Society, 31 (6), 607–630. https://doi.org/10.1177/0957926520939690

Lee, E.-J. (2007). Deindividuation Effects on Group Polarization in Computer-Mediated Communication : The Role of Group Identification, Public-Self-Awareness, and Perceived Argument Quality. Journal of Communication, 57(2), 385- 403. https://doi.org/10.1111/j.1460-2466.2007.00348.x

Massanari, A. (2017). #Gamergate and The Fappening : How Reddit's algorithm, governance, and culture support toxic technocultures. New Media & Society, 19(3), 329- 346. https://doi.org/10.1177/1461444815608807

Munn, L. (2019). Alt-right pipeline : Individual journeys to extremism online. First Monday. https://doi.org/10.5210/fm.v24i6.10108

Partington, Alan, éd. 2013. Patterns and meanings in discourse: theory and practice in corpus-assisted discourse studies (CADS). Amsterdam: John Benjamins Publishing Company.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wright, S., Trott, V., & Jones, C. (2020). ‘the pussy ain’t worth it, bro’: Assessing the discourse and structure of MGTOW [Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2020.1 Information, Communication & Society, 23 (6), 908–925. https : / / doi . org / 10 . 1080 / 1369118X.2020.1751867

L'interprétation des jeux de mots sur le réseau social X

LIU Haoran

Centre de Recherche sur les Médiations,
Université de Lorraine, DAMAS Laboratory, 57045 Metz, France

liuhaoran199412@gmail.com

Mots-Clés : réseau social X ; jeu de mots ; interprétation

Doctorant en sciences du langage en rédaction d'une thèse au sujet linguistique basée sur un corpus numérique, je proposerais une présentation qui porte sur l'analyse des jeux de mots sur le réseau social X.

Le jeu de mots est un phénomène dans lequel un locuteur produit un énoncé qui juxtapose ou manipule les éléments linguistiques d'une langue ou de plusieurs langues afin de surprendre les auditeurs et provoquer un effet humoristique (Winter-Froemel 2016). Le jeu de mots en tant que sujet de recherche est beaucoup développé par les auteurs différents. Freud (1905) définit le jeu de mots comme le jeu sur le double sens d'un mot et il classifie le jeu de mots comme un sous-type du mot d'esprit. Pierre Guiraud (1979) définit le jeu de mots comme « les jeux sur les mots » et il distingue le jeu de mots du mot d'esprit et aussi du divertissement verbal. En 2015, le premier volume de Dynamics of Wordplay est sorti. Cette série du livre éditées par Winter-Froemel englobe les approches différentes sur le jeu de mots dans les dimensions linguistiques, littéraires et pragmatiques chez les chercheurs différents. Ce résumé des recherches sur le jeu de mots depuis 1905 remontre que la recherche sur le jeu de mots ne se limite plus à la définition du jeu de mots, la délimitation entre le jeu de mots et les autres types du jeu verbal et les techniques de la production des jeux de mots. La recherche sur l'utilisation du jeu de mots dans la dimension pragmatique est devenue le sujet de tendance. X en tant que réseau social propose une occasion à ces utilisateurs de communiquer entre eux. Cette communication sur les réseaux sociaux implique l'utilisation pragmatique des jeux de mots.

Le réseau X (anciennement Twitter) est le plus symbolique de nouvelles formes d'expression en ligne. Twitter est fondé sur une expression essentiellement publique qui est accessible à tous sans filtre d'abonnements ou de liens d'amitiés et il permet l'expression de tout un chacun sur une toile interconnectée dans le monde entier et ouvre la possibilité d'un espace public sans limite (Douay, Reys et Robin, 2015 : 29). En raison des fonctionnements de X tel que l'utilisation des images et vidéos, la datation d'un tweet, l'utilisation de hashtags et les retweets, les éléments dont les auditeurs ont besoin pour reconnaître, comprendre et interpréter les jeux de mots sont très divers.

Dans le but de comprendre comment les utilisateurs utilisent les jeux de mots pour communiquer entre eux, j'ai choisi d'effectuer ma recherche sur les techniques de la production des jeux de mots, l'interprétation des jeux de mots dans la communication interactive sur X et aussi les fonctions des jeux de mots sur X en analysant les captures d'écrans des tweets qui contiennent des jeux de mots comme corpus.

L'aspect essentiel sur la méthodologie de ma recherche est la construction de mon corpus. Afin d'éviter les comptes humoristes dans la mesure du possible, j'ai choisi de de m'abonner des comptes sur X d'une façon semi-aléatoire. Voici la méthode précisée : j'ai choisi de m'abonne aux 20 comptes qui postent un tweet sur le premier thème dans Tendances : France le premier et le 16 du mois, c'est-à-dire, deux fois par mois. L'abonnement avait commencé au premier mars 2020 et a terminé au premier août 2021. L'abonnement aux 700 comptes me permet de construire mon corpus. D'après ma recherche sur les exemples du jeu de mots dans mon corpus, je présente les premiers résultats de ma recherche.

Les techniques de la production des jeux de mots peuvent être divisées en deux catégories: les techniques au niveau sub-lexical qui contiennent le jeu sur les éléments phonétiques, morphologiques, syllabiques et orthographiques. Les techniques au niveau lexical contiennent l'homonymie ou polysémie, la paronomase, le jeu sur les éléments phraséologiques, le jeu sur lexical set, le jeu sur les relations formelles, l'opposition sémantique et le jeu de mots syntaxique. En outre, les fonctionnements de X proposent aussi des techniques particulières pour produire les jeux de mots (ex. Les locuteurs peuvent utiliser des emojis pour produire des jeux de mots.)

Les jeux de mots sur X peuvent avoir les fonctions linguistiques et les fonctions pragmatiques. Ils peuvent avoir potentiellement trois fonctions linguistiques : la fonction métalinguistique, la fonction poétique et la fonction expressive. Ils peuvent aussi avoir les fonctions pragmatiques selon les discours. Par exemple, les jeux de mots sur X peuvent avoir une fonction sociale (Winter-Froemel, 2016 : 37) de montrer l'image du soi de locuteur (ex. Les comptes humoristiques postent des jeux de mots pour partager le plaisir esthétique avec les autres utilisateurs et montrer leur maîtrise du langage, etc.) Les jeux de mots sur X peuvent souvent produire un effet humoristique mais parfois l'humour peut être absent dans certaines situations surtout dans les tweets avec un sujet sérieux tel que la guerre. Les jeux de mots peuvent avoir d'autres fonctions pragmatiques (ex. le locuteur utilise des jeux de mots dans ses tweets politiques pour se moquer des hommes politiques, etc.)

Pour interpréter les jeux de mots, l'auditeur a besoin d'éléments divers. Les compétences linguistiques (les compétences grammaticales, lexicales, phonétiques, syntaxiques, etc.) sont

des éléments indispensables dans tous les jeux de mots. Les dispositifs de X (les fonctionnements de X tel que la datation, l'accompagnement de l'image et de la vidéo, etc.) sont très importants pour l'interprétation des jeux de mots (ex. Onze mai d'accord est un jeu de mots posté par un utilisateur le 11 mai). L'interprétation des jeux de mots implique aussi les informations hors des dispositifs de X tel que la connaissance personnelle, la situation politique, les éléments cultures, etc. (ex. Pour interpréter Pas saine !, il faut connaître que les internautes pensent que la Seine est trop polluée pour qu'on nage dedans.) Parfois, les trois éléments sont tous indispensables dans l'interprétation de certains jeux de mots. Cette question sur les éléments de l'interprétation peut se rapporter à la question de contexte.

Pour clairement présenter ma recherche lors de la communication, je propose un plan qui contient trois sous-parties. La première sous-partie porte sur une présentation de la construction de mon corpus. La deuxième sous-partie est une présentation générale sur les définitions et les techniques de jeu de mots. La troisième sous-partie montre les fonctions des jeux de mots sur X et les éléments nécessaires proposés par X et hors de X pour interpréter les jeux de mots.

Bibliographie

[1] Douay, Reys et Robin, « L'usage de Twitter par les maires d'Île-de-France », Netcom [En ligne], 29-3/4 | 2015, mis en ligne le 20 mai 2016, consulté le 16 mai 2024. URL: http://journals.openedition.org.bases-doc.univ-lorraine.fr/netcom/2089; DOI: https://doi-org.bases-doc.univ-lorraine.fr/10.4000/netcom.2089

[2] Freud, Le mot d'esprit et ses rapports avec l'inconscient (1905). Traduit de l'allemand par Marie Bonaparte et le Dr. M. Nathan en 1930. Paris: Gallimard, 1930. Réimpression : Gallimard, 1971, 378 pp. Collection idées, nrf, n 198.

[3] Guiraud, Les jeux de mots. Paris, Presses Universitaires de France, 1976.

[4] Jakobson, « Linguistique et poétique » dans Essaie de linguistique générale, 1963.

[5] Winter-Froemel, « Approaching Wordplay », Crossing Languages to Play with Words: Multidisciplinary Perspectives. Berlin, Boston : De Gruyter, p.37, 2016.https://doi.org/10.1515/9783110465600

[6] Winter-Froemel et Zirke. Enjeux du jeu de mots : Perspectives linguistiques et littéraires, Berlin, München, Boston : De Gruyter, 2015. DOI : https://doi.org/10.1515/9783110408348

Étude des langues fictives : une perspective linguistique dans une ère nouvelle

BALTACHE Imane

Université de Chlef Hassiba Benbouali, Algérie

Laboratoire technologies de l'Information et de la Communication dans l'Enseignement des Langues Etrangères et Traduction, Université de Chlef

i.baltache94@univ-chlef.dz

Mots-Clés : Langues fictives, typologie linguistique, linguistique computationnelle, création linguistique.

L'étude des langues fictives, longtemps perçue comme une curiosité, suscite aujourd'hui un intérêt grandissant au sein des milieux académiques, notamment en linguistique. Ces langues, créées dans des contextes littéraires, cinématographiques ou ludiques, représentent non seulement des défis de créativité, mais aussi des objets d'étude pertinents pour l'analyse linguistique. Les langues telles que le Klingon de Star Trek ou le Dothraki de Games of Thrones offrent des exemples concrets de systèmes linguistiques entièrement construits qui peuvent être étudiés pour comprendre les processus sous-jacents à la création et à l'évolution des langues. D'ailleurs Peterson (2015 :19), le créateur des langues comme le dothraki pour Games of Thrones, explique dans son ouvrage intitulé *The Art of Language Invention* que la création de langues fictives n'est pas simplement un exercice de créativité. Elle nécessite une compréhension approfondie de la linguistique pour que les langues soient à la fois plausibles et apprenables.

L'objectif de cette étude est de proposer une analyse structurale et typologique des langues fictives et d'explorer comment les technologies linguistiques computationnelles peuvent être appliquées à leur étude. En effet, ces langues permettent d'explorer des modèles linguistiques alternatifs et de tester les limites des théories linguistiques existantes. Cette recherche vise à combler le fossé entre la linguistique théorique et appliquée en intégrant des méthodes d'analyse moderne issues du domaine de l'intelligence artificielle et du traitement automatique des langues (TAL).

Afin de guider cette étude, plusieurs questions de recherche ont été formulées pour explorer en profondeur les aspects structuraux, typologiques et technologiques des langues fictives :

- Quels sont les traits structurales et typologiques communs que l'on retrouve dans les langues fictives ? Cette question vise à identifier les caractéristiques

phonologiques, morphologiques, syntaxiques et sémantiques récurrentes dans ses langues.

- Comment les outils et les techniques de la linguistique computationnelle peuvent-ils être adaptés pour analyser des langues fictives ? Ici, l'objectif est de déterminer comment les technologies actuelles, comme l'analyse syntaxique automatisée ou les modèles de langage, peuvent être appliquées à l'étude des langues fictives.
- Quels enseignements peut-on tirer de l'étude des langues fictives pour les compréhensions des processus cognitifs liés à la création et à l'apprentissage des langues ? Cette question examine les implications de la création des langues fictives sur notre compréhension de la faculté linguistiques humaine et des capacités d'innovation linguistique.

Pour répondre à ces questions, une approche combinant méthodes descriptives et expérimentales a été adoptée, intégrant des techniques traditionnelles de la linguistique et des outils de linguistique computationnelle.

En effet, l'approche adoptée dans cette étude est à la fois descriptive et expérimentale, combinant des méthodes traditionnelles de la linguistique et des techniques issues de la linguistique computationnelle.

- Analyse structurale et typologique : Une première phase descriptive consiste à cataloguer et à analyser un corpus de langues fictives en fonction de leurs traits structuraux. Cette étape s'appuie sur les méthodes de typologie linguistique pour comparer les langues fictives entre elles ainsi qu'avec des langues naturelles.
- Linguistique computationnelle : La seconde phase de l'étude utilise des outils de traitement automatique des langues pour modéliser et analyser les langues fictives. Cela inclut la création de grammaires génératives et l'entrainement de modèles de langage spécifiques à ces langues à l'aide d'algorithmes d'apprentissage automatique. Les techniques telles que parsing [5] ou la génération automatique de phrases seront employées pour tester la cohérence et la complexité des langues étudiées.
- Expérimentation cognitive : Enfin, des expériences cognitives pourraient être menées pour évaluer comment les individus perçoivent et apprennent ces langues fictives par rapport aux langues naturelles. Ces expériences viseraient à explorer la charge cognitive associée à l'apprentissage de langues construites, ainsi que leur potentiel pour révéler des aspects universels de la cognition linguistique.

[5] Analyse syntaxique

Les données pour cette étude proviendront d'un corpus établi de langues fictives, comprenant des manuels de grammaire, des textes originaux, des ressources en ligne. Un échantillon représentatif de langues fictives sera sélectionné en fonction de critères tels que la complexité grammaticale, la popularité et la documentation disponible.

Ainsi, les résultats attendus de cette recherche devraient révéler des schémas communs dans la structure des langues fictives, malgré leur diversité apparente. Par exemple, on s'attend à identifier des tendances universelles dans la construction des phonèmes ou des structure syntaxiques, ce qui pourrait renforcer ou contester certaines théories existantes sur les universaux linguistiques.

En ce qui concerne l'application des technologies computationnelles, on anticipe des défis techniques, mais aussi des succès significatifs, comme la mise au point de modèles capables de générer du texte dans une langue fictive avec une précision élevée. Les résultats pourraient ainsi ouvrir la voie à de nouvelles méthodes pour la création et l'analyse de langues construites, et offrir des perspectives innovantes pour l'étude des processus linguistiques.

En conclusion, cette étude se situe à l'intersection de la linguistique théorique, de la linguistique appliquée et de la technologie. En s'intéressant aux langues fictives, elle non seulement enrichit notre compréhension des potentialités linguistiques humaines mais contribue également à l'innovation dans le domaine du traitement automatique des langues. Les résultats obtenus devraient permettre d'affiner les théories linguistiques existantes et de développer de nouveaux outils pour l'analyse linguistique, tout en offrant des perspectives inédites sur la créativité linguistique.

Bibliographie

Adams, Michael, (2011), *From Elvish to Klingon: Exploring Invented Languages*. Oxford University Press

Comrie, Bernard, (1989). *Language Universals and Linguistic Typology*. University of Chicago Press.

Greenberg, Joseph H. (1963). *Universals of language*. MIT press.

Jurafsky, D., and Martin, J-H., (2023), *Speech and Language Processing (3rd edition)*, Pearson.

Manning, C., and Schutze, H., (1999), *Foundation of Statistical Natural Language Processing*.

Okrent, Arika, (2009), *In the Land of Invented languages: Adventures in Linguistic Creativity, Madness, and Genius*. Spiegel & Grau.

Peterson, David J., (2015), *The Art of Language Invention: From Horse-Lords to Dark Elves, the Words Behind World-Building*. Penguin Books.

Proposition d'un cadre d'analyse pour annoter l'évènement de la guerre en Ukraine dans les éditoriaux français, anglais et allemands

VERSMESSEN Marie

MoDyCo, Université Paris-Nanterre

marie.versmessen@gmail.com

Mots-Clés : analyse de discours contrastive, éditorial, guerre en Ukraine, modalité

Cette communication s'inscrit dans le champ de l'Analyse du Discours Contrastive (Claudel *et al.*, 2013 ; Pordeus Ribeiro, 2015 ; von Münchow, 2021). Elle a pour objectif de mettre au jour le cadre d'analyse d'une étude comparative portant sur des éditoriaux de la presse écrite française anglaise et allemande, traitant d'un évènement, la guerre en Ukraine (2022-). Pour répondre à cet objectif, on présentera tout d'abord la démarche méthodologique qui a conduit à élaborer une passerelle trans-langagière entre les trois langues et le schéma d'analyse qui en résulte. Ensuite, des résultats d'une analyse comparative menée sur un échantillon des corpus seront exposés. Cela nous conduira, dans un troisième temps, à dresser des hypothèses relatives aux caractéristiques langagières du genre et aux représentations de l'évènement transmises dans les journaux des trois pays.

La démarche méthodologique de cette étude a consisté à définir un mode d'articulation des corpus et de l'entrée langagière au sein de ces derniers.

Les corpus ont été rapprochés selon les critères de référentialité et d'orientation politique des organes de presse de chaque pays. Ils ont été ensuite recueillis sur une période identique. C'est ainsi que 193 éditoriaux traitant de la guerre en Ukraine, issus des trois grands organes de presse de chaque pays, *Le Monde*, *Libération*, *Le Figaro* ; *The Guardian*, *The Independent* et *The Times* ; et *Die Zeit*, *Die TAZ*, *Die Frankfurter Allgemeine Zeitung* ont été collectés sur la période de février-mars 2022, marquant le début (officiel) de la guerre menée par la Russie en Ukraine.

L'entrée d'analyse a été choisie à la suite d'une étude portant sur les caractéristiques du genre éditorial dans les travaux en science de l'information et de la communication (notamment Agnès, 2015 ; Firmstone, 2019 ; Mast, 2018) et de lectures exploratoires de nos corpus. Les éditoriaux français, anglais et allemands transmettant l'opinion du journal, le choix s'est porté sur une opération discursive intervenant dans les segments de jugement et d'appréciation : la

modalité. Pour élaborer une passerelle trans-langagière dans nos corpus au travers de cette notion, on s'est interrogé sur la façon dont la notion se conçoit et se structure dans les trois langues. Le constat auquel nous sommes arrivée est que la modalité, quel que soit le système linguistique auquel elle est rattachée, est une notion relativiste. Celle-ci pouvant en effet intervenir à tous les niveaux de la langue (morphologique, syntaxique, lexical, etc.), les linguistes la restreignent en fonction des objectifs fixés à leur étude. A titre d'exemples, la modalité en français est adoptée différemment chez Gosselin (2010) et Galatanu (2005). Alors que Gosselin l'envisage à l'échelle des lexèmes et des grammèmes, Galatanu, elle, se concentre sur les marqueurs modaux lexicaux et leurs enchaînements discursifs argumentatifs. De même, la modalité en anglais conçue chez Palmer (2001) comme une catégorie grammaticale relativement unifiée, apparaît chez Nuyts (2006) tout au plus comme un concept heuristique comparatif en raison de l'ensemble des opérations linguistiques et sémantiques auxquelles elle renvoie. La modalité en allemand est également traitée différemment chez Rousseau (2003) qui préfère se concentrer sur l'étude des verbes modaux pour réduire les possibilités de fluctuation et chez Modicom (2014) qui focalise son étude sur les particules modales.

Ces approches relativistes posent souci à l'entreprise comparative car elles ne permettent pas aisément de dresser un aperçu exhaustif de la modalité en français, en anglais et en allemand et donc de poser la comparabilité de la notion. Face à cette difficulté, une passerelle trans-langagière a été élaborée à un niveau extérieur des systèmes linguistiques, c'est-à-dire au niveau sémantique des valeurs modales que les trois langues partagent : les valeurs ontologiques, de jugements de vérité, axiologiques et finalisantes.

Afin de décrire en discours ces valeurs modales qui relèvent des jugements, des attitudes et des appréciations du sujet parlant, une approche onomasiologique a été adoptée. Elle consiste tout d'abord à donner à chaque valeur modale, une définition solide et opérationnelle à partir de laquelle il est possible d'identifier les constructions qui relèvent de la valeur modale en question. Il s'agit ensuite de décrire ces constructions à l'aide d'un outil d'annotation manuelle, Analec (Landragin et al., 2012), à un niveau conceptuel, avec un paramètre sémantique rendant compte de la valeur modale précise du marqueur ; à un niveau structurel avec un paramètre syntaxique décrivant la forme de la modalité, marquée linguistiquement ou inférée par le discours ; à un niveau énonciatif explicitant les degrés de prise en charge ; et à un niveau discursif avec un paramètre lié à la fonction descriptive ou injonctive du marqueur. On fait alors l'hypothèse que cette approche descriptive nous permette l'accès aux caractéristiques génériques de l'éditorial et aux représentations de l'évènement dans les journaux des trois pays.

Une première analyse qualitative portant sur trois éditoriaux issus de chacun des journaux suivants *Libération, The Independent* et *Die TAZ* a révélé la prévalence de la modalité axiologique (marquée et inférée). Les marqueurs axiologiques ne se contentent pas seulement de décrire l'évènement mais donnent une image (l'*éthos* de la rhétorique aristotélicienne) des actants engagés dans la guerre. Ils rendent également possible, pour certains d'entre eux, la persuasion orientée vers l'action en fournissant des raisons d'agir. Le poster illustrera ces premières observations.

Bibliographie

AGNES, Y. (2015) : *Manuel de journalisme*, La Découverte, Paris

CLAUDEL, Ch., VON MUNCHOW, P., PORDEUS RIBEIRO M., PUGNIERE-SAAVEDRA, F., TREGUER-FELTEN, G. (dir.) (2013) : *Cultures, discours, langues, Nouveaux abordages*, Limoges, Lambert-Lucas

DENDALE, P. & COLTIER, D. (2005). « La notion de prise en charge ou de responsabilité dans la théorie scandinave de la polyphonie linguistique », in : J. Bres *et al*. (éd.), pp. 125-140

FIRMSTONE, J. (2019) : *«* Editorial journalism and newspapers' editorial opinions *»*, in : *Oxford Research Encyclopedia of Communication,* Oxford University Press, URL : https://www.researchgate.net/publication/342529445_Firmstone_J_2019_Editorial_journalism_and_newspapers'_editorial_opinions_In_Oxford_Research_Encyclopedia_of_Communication_Oxford_Research_Encyclopedia_Oxford_University_Press_Oxford_England [consulté 13/09/2024]

GALATANU, O. (2005) : « La sémantique des modalités et ses enjeux théoriques et épistémologiques dans l'analyse des textes », in : J.-M. Gouvard (éd.), pp. 157-170

GOSSELIN, L. (2010) : *Les modalités en français*, Amsterdam-New York, Rodopi

KERBRAT-ORECCHIONI, C. (2002) : *L'énonciation*, Paris, Armand Colin

LANDRAGIN, F., POIBEAU, T. et VICTORRI, B. (2012) : « Analec : a new tool for the dynamic annotation of textual data », in : *International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, URL: https://shs.hal.science/halshs-00698971v1/document [consulté le 13/09/24]

MAST, C. (2018) : *ABC des Journalismus, Ein Handbuch*, Herbert von Halem Verlag, Köln

MODICOM, P.-Y. (2021) : « Les théories de la prise en charge au prisme des particules modales de l'allemand », Le sens entre langue et discours : études de sémantique et d'analyse de discours, ELIS, pp. 61-80 pp. 47-69, URL : https://shs.hal.science/halshs-01090454v1 [consulté le 13/09/2024]

NISSIM, M., PIETRANDREA, P. (2017) : « Modal : a multilingual corpus annotated for modality », in : R. Basili, M. Nissim & G. Satta (éd.), *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017*, Academia University Press, URL: https://doi.org/10.4000/books.aaccademia.2435 [consulté le 13/09/24]

NUYTS, J. (2006) : « Modality: Overview and Linguistic Issues », in : Frawley, W., *The Expression of Modality*, Berlin, Mouton de Gruyter, pp. 1-26

ROUSSEAU, A. (2003) : « La question des verbes de modalité en allemand », *Langues et littératures modernes*, pp. 797-823

VON MUNCHOW, P. (2021) : *L'analyse du discours contrastive. Théorie, méthodologie, pratique*, Limoges, Lambert-Lucas

Innovations pédagogiques du projet écri+ : perspectives de recherche

LE COZ DENTU Fanny[1,2,3] et GAUDRAY BOUJU Vanessa[2,3]

1 PREFICS ; 2 MoDyCo ; 3 écri+

fanny.le-coz-dentu@univ-rennes2.fr ; v.gaudraybouju@gmail.com

Mots-Clés : écri+, français académique, réussite étudiante, didactique du français, français langue étrangère, français sur objectif universitaire, TAL, reprises

Le projet ANR écri+ est né en 2018 du constat de la difficulté de maîtrise du français académique chez les étudiants de l'enseignement supérieur. Cette difficulté s'explique par un double mouvement : l'université se démocratise depuis les années 1980 tandis qu'en parallèle les programmes scolaires se sont densifiés, laissant moins de temps à l'apprentissage de la langue écrite (Boch et Buson, 2012). D'autre part, les types de discours et les attentes des enseignants du supérieur en termes de compétences en lecture, capacité de synthèse ou encore qualité de rédaction sont nouveaux pour les étudiants entrant à l'université (Frier, 2020) et constituent ce qu'on appelle un « événement littéracique majeur » (Jaffré, 2004).

Pour aider les étudiants à développer leurs compétences en français écrit, écri+ bénéficie d'un partenariat avec plusieurs universités et spécialistes francophones issus de différents domaines (linguistique, informatique, didactique, etc.) qui contribuent à l'élaboration de dispositifs de formation, parmi lesquels :

- Un référentiel de compétences précis spécifique au français écrit, prenant en compte toutes ses dimensions (orthographe, lexique, syntaxe, compréhension et structuration de texte, registres de langue, prise en compte des points de vue, etc.), construit à partir d'erreurs avérées et recueillies dans des écrits d'étudiants (De Luca & al., 2022) ;
- Une plateforme de formation en ligne (*écri+test*) permettant aux étudiants de travailler des compétences rédactionnelles en autonomie au travers d'exercices adaptés du référentiel ;
- La possibilité pour l'utilisateur d'*écri+test* de faire certifier le niveau atteint au moyen d'un examen, également basé sur le référentiel de compétences.

Nous présentons ici deux perspectives de recherche, abordées dans deux thèses distinctes commencées en 2023, qui visent respectivement à élargir le public ciblé (étudiants internationaux) et les outils à disposition (détection automatique d'erreurs).

écri+, outil de FLE/FOU ? (Thèse de Fanny Le Coz Dentu)

L'une de ces nouvelles perspectives de recherche découle du constat suivant : les outils décrits ci-dessus sont utilisés à la fois par un public d'étudiants francophones natifs (FLM) mais aussi par un public d'apprenants ayant le français comme langue étrangère (FLE), alors que le dispositif n'a pas été pensé pour répondre aux spécificités de ce type d'apprenant. Ce dernier public comprend notamment des personnes qui souhaitent poursuivre des études supérieures en langue française et dont les besoins peuvent relever du Français sur Objectif Universitaire (FOU) (voir Bordo & al., 2016).

C'est pourquoi il est ici question de déterminer la pertinence des outils développés par écri+ pour le public des étudiants étrangers potentiels ou actuels. Ainsi, il permettrait d'examiner la possibilité d'intégrer les outils de formation et de certification écri+ aux dispositifs de FLE/FOU proposés à ces étudiants, voire de les adapter à ce public particulier. Plusieurs axes de travail sont proposés pour répondre à cette problématique.

Tout d'abord, une analyse des difficultés en français écrit rencontrées par les étudiants étrangers de niveau intermédiaire à avancé (B2-C2 du CECRL (2001)) servira à évaluer leurs spécificités par rapport à celles rencontrées par les étudiants natifs du français (voir Lang, 2019). Pour cela, un corpus très divers de copies d'examen d'étudiants de Français Langue Étrangère a été recueilli (écrits argumentés, écrits créatifs, restitution de cours…) et permet d'étudier la plupart des types de difficultés répertoriées par le référentiel écri+.

Ensuite, nous mesurerons à l'aide d'outils statistiques la progression d'étudiants francophones natifs et d'étudiants ayant appris le français plus tardivement suivant le même programme d'entraînement sur *écri+test*. Ces étudiants, les première année de Licence de l'Université de Rennes 2 (toutes disciplines confondues), seront également interrogés sur la perception de leurs propres difficultés. Les premiers résultats montrent que les étudiants FLE progressent au moins autant que les FLM (selon le prisme écri+) ; en revanche, la sensation de maîtrise des FLM des différentes compétences diminue entre le début et la fin du programme de formation, tandis que celui des FLE tend plutôt à augmenter.

Enfin, des tests d'*écri+test* et de ses fonctionnalités avec des étudiants non francophones natifs au moyen d'enquêtes par questionnaire complétées par des entretiens seront organisés afin de donner la parole à ce nouveau public cible, recueillir ses besoins en matière de maîtrise du français écrit académique, et mieux envisager la capacité du projet écri+ à y répondre.

Vers un outil de détection d'erreurs de reprise (Thèse de Vanessa Gaudray Bouju)

Pour compléter sa palette de ressources, écri+ développe son propre outil d'aide à la correction, adapté aux exigences de l'université. Certains correcteurs (ProLexis, Antidote, Cordial) permettent déjà de repérer certaines erreurs (orthographe, grammaire), mais d'autres sont plus difficiles à détecter automatiquement, comme celles concernant la cohérence du discours, qui requièrent une compréhension plus fine de la langue et du contexte. Or, la détection de ces phénomènes est un enjeu pour le projet écri+ : un outil capable de repérer les erreurs ou maladresses que font les étudiants dans leurs textes afin de les leur signaler les aiderait à être plus indépendants dans l'analyse et la correction de leurs écrits.

La thèse de Noreskal (2022) s'est focalisée sur un cas spécifique, les structures coordonnées erronées, afin de mettre au point un outil les détectant automatiquement dans des rédactions étudiantes. Il est question, à travers cette nouvelle thèse qui a débuté en 2023, de poursuivre le développement de cet outil en l'étendant aux problèmes de reprises.

Les reprises, notion englobant celles d'anaphore et de coréférence, consistent en la mention d'un élément dont l'antécédent a déjà été introduit dans le discours. Il est nécessaire de pouvoir retracer le lien de référence entre les deux entités pour que la cohésion du texte soit assurée. Les reprises aident en effet à structurer le discours en maintenant ou en faisant évoluer un thème. Néanmoins, leur maîtrise n'est pas toujours aisée. Reichler-Béguelin (1988) a par exemple montré que certaines erreurs sont dues au fait qu'un élément peut être saillant dans l'esprit du locuteur mais pas dans le discours, ce qui entrave son interprétation. Schnedecker (1995) a pour sa part insisté sur le rôle de l'enseignement des reprises au primaire, qui se focalise sur le fait « d'éviter les répétitions » mais cause en contrepartie des problèmes de continuité référentielle dans des textes d'élèves cherchant à suivre cette règle.

Un corpus de rédactions étudiantes (issues de cursus variés allant de la L1 au doctorat) est en cours de constitution pour inventorier et analyser en détail les difficultés liées à l'usage des reprises. Les premières analyses des textes recueillis ont fait apparaître différents types de problématiques : erreurs grammaticales, ambiguïtés référentielles ou encore maladresses stylistiques. Celles-ci font l'objet de sanction par les correcteurs car elles entament la fluidité du texte, voire gênent sa compréhension. Néanmoins, la divergence des dysfonctionnements relevés va nous amener à questionner le statut d'*erreur* à travers les notions d'*acceptabilité* et de *gravité*. À partir de nos analyses et de la typologie des erreurs de reprise qui sera constituée, nous pourrons entraîner, grâce aux techniques de TAL modernes (apprentissage profond), un outil d'aide à la correction innovant adapté aux besoins des étudiants.)

Bibliographie

Boch, F., & Buson, L. (2012). Orthographe & grammaire à l'université. Quels besoins? Quelles démarches pédagogiques? *Scripta*, 16(30), Article 30.

Bordo, W., Goes, J., Mangiante, J.-M. (Éds.) (2016). Le Français sur objectif universitaire : Entre apports théoriques et pratiques de terrain. *Artois Presses Université.* https://books.openedition.org/apu/13673

Conseil de l'Europe. (2001). Cadre européen commun de référence pour les langues (CECRL). éduscol | Ministère de l'Éducation nationale et de la Jeunesse - Direction générale de l'enseignement scolaire.

De Luca, G., De Vogue, S., Lefebvre, J., & Sitri, F. (2022). écri+, un dispositif en ligne d'évaluation, de formation et de certification des compétences écrites en français : Le cas de la citation, entre formation et recherche. TEISEL. *Tecnologías para la investigación en segundas lenguas*, 1. https://doi.org/10.1344/teisel.v1.37075

Frier, C. (2020). Les défis de l'enseignement supérieur et l'état des recherches sur les littéracies universitaires. Écrire dans l'enseignement supérieur.

Jaffré, J.-P. (2004). La litéracie : Histoire d'un mot, effets d'une notion. La littéracie. *Conceptions théoriques et pratiques d'enseignement de la lecture-écriture*, 21. https://shs.hal.science/halshs00089961

Lang, E. (2019). L'écrit(ure) universitaire, une tâche située et complexe : approche holiste du processus d'adaptation de la compétence scripturale chez les apprenants avancés en FLE. Linguistique. Université de Strasbourg. Français. NNT : 2019STRAC024. tel-03018108

Noreskal, L. (2022). Erreurs dans les phrases coordonnées au sein des rédactions universitaires : Typologie et détection [These de doctorat, Paris 10]. https://theses.fr/2022PA100148

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.

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.

Etude morphologique des suffixes -ance et -ence en français

LIN Yifeng

Université Paris Nanterre, MoDyCo - CNRS (UMR 7114)

LYFLinguistique@163.com

Mots-Clés : vocabulaire ; l'ère numérique ; manuel ; FLE

Cette recherche porte sur la description des propriétés morphologiques des suffixes *-ance* et *-ence*. Les suffixations en *-ance*/*-ence* jouent un rôle important dans le phénomène de nominalisation en français. Elles ont tendance à sélectionner une base verbale (ACCOINTANCE < ACCOINTER ; FLOCULENCE < FLOCULER) ou les bases adjectivales en *-ant* (CLAIRVOYANCE < CLAIRVOYANT) et *-ent* (INDULGENCE < INDULGENT), et ont récemment fait l'objet de nombreux travaux de recherche (Dal & Namer 2010 ; Knittel 2016 ; Knittel & Marin 2021 ; Gonzalez 2022 ; Gréa et al. 2023). La majorité de ces recherches se concentrent sur la base sélectionnée en sémantique et sur la structure morphologique de la base, mais peu se sont intéressées aux autres propriétés morphologiques de la base, telles que la productivité morphologique, le groupe de conjugaison de la base verbale ainsi que la combinaison suffixale. Pourtant, ces trois propriétés morphologiques occupent une place primordiale en morphologie et déterminent souvent le choix du schéma morphologique adopté pour dériver les lexèmes (Hilpert 2013 ; Anscombre 2015 ; Aronoff 2019). A titre d'exemple, à travers l'exploitation de niches orthographiques et de la combinaison suffixale dans les suffixations en *-ance*/*-ence*, Aronoff (2019) a montré que le suffixe *-ance* en anglais se distinguait du suffixe *-ence* selon ces deux aspects et a proposé que la niche orthographique du radical permettait d'être discriminante pour les suffixes en compétition (dans de nombreuses recherches menées, le suffixe *-ence* est considéré comme une variante orthographique du suffixe *-ance*.)

À partir de ces travaux de recherche, les problématiques de notre étude sont les suivantes : 1) Quelle est la productivité diachronique des suffixes *-ance*/*-ence* ? 2) Quel(s) sont le(s) groupe(s) de conjugaison du verbe préférentiel(s) pour les deux suffixes ? 3) A l'aide de niches orthographiques et de combinaisons suffixales, le suffixe *-ance* se distingue-t-il à l'écrit du suffixe *-ence* pour le français, sur le modèle de ce qui a été observé en anglais (Aronoff 2019) ?

Notre étude se fonde sur des données dictionnairiques en provenance du *Grand Robert de la langue française* (2005). Pour annoter ces données, nous avons adapté la méthodologie

proposée par Dal & Namer (2010) : 1) analysabilité[6] des adjectifs de la forme X*ant*/X*ent* où X est un verbe ; 2) analysabilité des adjectifs de la forme X*ant*/X*ent* dépourvus de verbes correspondants, au moins en synchronie ; 3) analysabilité sans médiation par une étape adjectivale ; 4) analysabilité des adjectifs de la forme X*ant*/X*ent* où X est un nom ; 5) analysabilité sans médiation par une étape adjectivale. Comme le montre le Tableau 1, nous avons retenu 238 noms en -*ance* et 163 noms en -*ence*.

Tableau 1 Répartition des noms en -*ance*/-*ence* selon six cas

	-*ance*	-*ence*
X-*ance*/-*ence* < X-*ant*/-*ent* < X_V	128	40
X-*ance*/-*ence* < X-*ant*/-*ent*	43	92
X-*ance*/-*ence* < X_V	52	9
X-*ance*/-*ence* < X-*ant*/-*ent* < X_N	0	13
X-*ance*/-*ence* < X_N	15	3
Autre	0	6

Lors de la présentation de nos résultats, nous allons avant tout nous pencher sur la productivité diachronique des deux suffixes. La Figure 1 montre que le suffixe -*ance* est généralement plus productif que le suffixe -*ence*, sauf au 14ème et au 19ème siècles, où leur productivité était similaire, et au 18ème siècle, où le suffixe -*ence* a connu un certain succès par rapport au suffixe -*ance*.

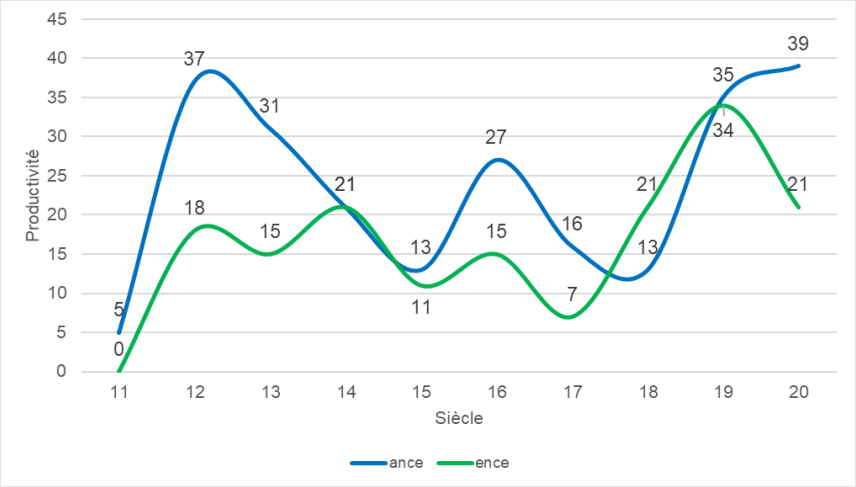

Figure 1 Productivité diachronique des suffixes -*ance* et –*ence*

[6] Par ce terme, Dal & Namer (2010) « [entendent] adopter un point de vue strictement synchronique de conformité aux patrons constructionnels que l'on peut dégager de l'observation des données actuelles, étant entendu que l'emprunt à une autre langue n'interdit pas l'analysabilité en français. »

Par ailleurs, pour ce qui est du groupe de conjugaison de la base verbale, comme illustré par le Tableau 2, les deux suffixes préfèrent le 1er et le 3ème groupe. Concrètement, le suffixe -*ence* présente une préférence pour des bases verbales du 1er groupe, quand le suffixe -*ance* prédomine de façon significative avec celles du 3ème groupe ($\chi^2 = 6.46$, df=1, p<0.05).

Tableau 2 Répartition des noms en -*ance*/-*ence* selon le groupe de conjugaison des bases verbales

	-*ance*	-*ence*
1er groupe	113 (63%)	43 (86%)
2ème groupe	9 (5%)	0 (0%)
3ème groupe	58 (32%)	7 (14%)

Selon Bonami & Boyé (2003) et Lin (2024), il existe plusieurs sous-classes de conjugaison en français. Pour les bases verbales du 3ème groupe, nous nous appuyons sur ces deux recherches afin d'exploiter les sous-classes de conjugaison préférentielles des suffixes -*ance*/-*ence*. Le Tableau 3 montre que le suffixe -*ance* a tendance à sélectionner plus de sous- classes que le suffixe -*ence*, pour lequel les sous-classes *valoir* et *courir* sont spécifiques.

Tableau 3 Répartition des bases verbales des noms en -*ance*/-*ence* dans le 3ème groupe selon les sous-classes de conjugaison

Suffixes	Sous-classes de conjugaison
-*ance*	*tenir* (12) ; X*ire*, *plaire* (9) ; X*ître* (7) ; X*endre* (7) ; *voir* (5) ; *partir* (3) ; *mettre* (2) ; X*evoir* (2) ; *faire* (2) ; *seoir* (2) ; *croire* (1) ; *dormir* (1) ; *faillir* (1) ; *mouvoir* (1) ; *soudre* (1) ; X*ffrir* (1)
-*ence*	*tenir* (2) ; *valoir* (2) ; X*ître* (1) ; *courir* (1) ; *mettre* (1)

Il nous reste encore à traiter les noms en -*ance*/-*ence* avec les bases verbales du 1er groupe. Nous avons recours aux niches orthographiques des radicaux pour notre analyse du français écrit. Comme le Tableau 4 l'illustre, un des deux suffixes est sélectionné préférentiellement pour chaque niche orthographique (*par ex.* X-*fér*, X-*flu* et X-*erg* sélectionnent le suffixe - *ence*).

Tableau 4 Niches orthographiques productives des radicaux des noms en -*ance*/-*ence*

Suffixes	Niches orthographiques
-ance	X-*ist* (6) ; X-*ntr* (4) ; X-*ord* (3) ; X-*son* (3) ; X-*onn* (3) ; X-*min* (3) ; X-*mbl* (3)
-ence	X-*fér* (7) ; X-*flu* (6) ; X-*erg* (5) ; X-*hér* X- (2) ; X- *nér* (2) ; X-*cul* (2)

Pour terminer, nous nous sommes intéressés à la combinaison suffixale des deux schémas morphologiques. Nous avons pu noter que lorsque les suffixes *-ance*/*-ence* constituaient un suffixe interne, ils avaient une propension à sélectionner des suffixes différents comme le suffixe externe pour construire une combinaison suffixale. A titre d'exemple, le suffixe *-ance* préfère les suffixes *-ier*/*-ière* pour construire des lexèmes (*-ance* + *-ier* > *-ancier* : CORRESPONDANCIER). En revanche, la combinaison X-*entiel* est plus fréquente avec le suffixe *-ence* (*-ence* + *-el* > *-entiel* PREFERENTIEL).

En conclusion, à travers l'analyse des propriétés morphologiques des suffixes *-ance/-ence*, nous avons observé que le premier suffixe était généralement plus productif que le second et qu'ils étaient sélectionnés différemment en fonction du groupe de conjugaison des bases verbales, des niches orthographiques des radicaux et de la combinaison suffixale. Cela pourrait étayer, dans une certaine mesure, l'hypothèse d'Aronoff (2019) : en langue écrite, pour les suffixes *-ance* et *-ence*, « la langue fait la distinction entre les deux, malgré leur homophonie. »

Bibliographie

Anscombre, J-C. (2015). Pour une approche morphosémantique des noms d'action en -tion, -ment et -age. *Revue de sémantique et pragmatique. 35-36* : 49-80.

Aronoff, M. (1994). *Morphology by itself, stems and inflectional classes.* (Linguistic Inquiry Monograph 22.) Cambridge, MA: MIT Press.

Aronoff, M. (2019). Competitors and Alternants in Linguistic Morphology. In: Rainer, F., Gardani, F., Dressler, W., Luschützky, H. (eds) *Competition in Inflection and Word-Formation. Studies in Morphology*, vol 5. Springer, Cham.

Bonami O. & Boyé G. (2003). Supplétion et classes flexionnelles. *Langages*. 37(152) : :102-126.

Dal, G. & Namer, F. (2010). Les noms en -ance/-ence du français : quel(s) patron(s) constructionnel(s) ? *Actes du CMLF 2010, Paris: ILF*, 893–907.

Hilpert, M. (2013). *Constructional change in English: Developments in allomorphy, word formation, and syntax.* Cambridge/New York: Cambridge University Press.

Gonzalez, O. (2022). Les nominalisations en -ance en francais. *Actes du Congrès annuel de l'Association canadienne de linguistique 2022*.

Grand Robert de la langue française. (2005). Robert : Paris.

Gréa, Philippe et al. Innovative uses of French neological -ance nominalizations. *ISMo 2023*, Sep 2023, Nancy, France.

Knittel, M. (2016). Les noms en -ance: un panorama. *SHS Web of Conferences 27*.

Knittel, M. & Marin R. (2021). Developing a resource for ance nouns, and related verbs and adjectives. *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*: 86-94.

Lin, Y. (2024). Étude statistique multivariée sur la flexion des verbes français : le cas des verbes du troisième groupe. *Etude Francophone* 2024 (03) : 49-57.

Les défis d'exportation et de traitement d'un corpus Facebook en Sciences du Langage

AELENEI Andreea Ioana

Université d'Orléans, France / Université Alexandru Ioan Cuza de Iași, Roumanie

ioana_aelenei@yahoo.com

Mots-Clés : nouchi ; Côte d'Ivoire ; discours numérique ; réseaux sociaux ; Python

Grâce à leur utilisation de plus en plus massive, les réseaux sociaux constituent un nouveau terrain de recherche intéressant pour les linguistes. Cependant, l'exportation et le traitement des données provenant de ces plateformes soulèvent certains défis techniques, pour lesquels nous essayerons de proposer quelques solutions. La présente communication concerne en particulier la création et l'exploitation d'un corpus Facebook, qui se trouve à la base d'une recherche doctorale en cours. Tout d'abord, c'est la collecte des données Facebook qui peut se relever problématique lorsqu'on essaie de récupérer des publications avec tous les commentaires hiérarchisés, à cause de la structure plutôt complexe de la plateforme. Alors que les logiciels que nous avons testés n'aboutissent pas au résultat souhaité, la source HTML de la page n'affiche pas non plus l'ensemble du contenu visé. Nous allons montrer comment nous avons procédé, en combinant le traitement automatique et manuel des données, afin d'exporter un nombre assez élevé de commentaires provenant de plusieurs pages différentes, dans un format qui se prête à une exploitation linguistique.

La question qui se pose ensuite est celle de l'analyse d'un corpus de ce type. Le discours numérique en général, et surtout celui des réseaux sociaux, implique un certain degré de variation orthographique (Fairon et Klein, 2010). Dans les interactions virtuelles informelles se développe une écriture non standard, qui s'éloigne de la norme jusqu'à l'introduction des chiffres, voire des symboles non alphanumériques dans l'orthographe. En outre, il ne faut pas oublier le caractère « composite » (Paveau, 2017 : 28) du discours numérique, qui intègre souvent des émoticônes, stickers, images, etc., rendant indiscernable la frontière entre langagier et technologique. Ce type d'éléments peuvent rendre plus difficile l'exploitation d'un tel corpus à travers des outils classiques de traitement de texte comme TXM. Ainsi, nous avons dû opérer une série de modifications sur le corpus avant d'aboutir à une version exploitable.

En outre, notre corpus se focalise sur le nouchi, une variété de français de Côte d'Ivoire. Ce parler, à l'origine un argot d'Abidjan (années 1970), s'est vite répandu dans le pays et

représente de nos jours un symbole de l'identité nationale ivoirienne, car il se fonde sur les créations lexicales hybrides et sur les emprunts aux langues autochtones (Boutin et Kouadio, 2015). Alors que son premier vecteur d'expansion était la musique, avec l'essor des réseaux sociaux, le nouchi a trouvé un autre milieu propice pour se répandre. Mais ce qui soulève des défis d'analyse, c'est justement sa nature hybride et le fait qu'il est issu d'un discours par excellence oral et réfractaire à la norme du français standard. Sur Facebook et non seulement, les locuteurs du nouchi se confrontent à la nécessité de transposer leur discours à l'écrit, sans se rapporter à une norme orthographique quelconque – car même si les chercheurs y ont réfléchi (Ahua, 2010), les locuteurs n'en ont pas forcément connaissance. La variation orthographique qui en résulte, spécifique des langues non standardisées même en dehors du contexte des réseaux sociaux (Millour, 2020), peut relever d'autres problèmes pour les chercheurs qui essayent d'exploiter un corpus de ce type.

Nous allons donc proposer une éventuelle solution, intégrant, d'une part, le logiciel TXM, et d'autre part, le langage de programmation Python, utilisé parfois pour le traitement des données linguistiques (Hammond, 2020), et qui pourrait représenter une alternative dans l'analyse des corpus issus des réseaux sociaux. La première étape dans laquelle nous nous sommes servi de Python est celle de la collecte des données : ce langage de programmation a permis de « nettoyer » les pages enregistrées sous le format HTML et de regrouper les données dans des fichiers CSV. Ensuite, des scripts Python nous ont permis de rendre le corpus plus facilement exploitable. D'une part, nous avons remplacé les émoticônes par une série de termes les désignant, pour remédier au problème de leur affichage incomplet dans certains logiciels, dont TXM. D'autre part, toujours à l'aide de Python, nous avons remplacé tous les noms des utilisateurs Facebook par des pseudonymes. En effet, nous avons intégré dans le corpus uniquement des publications Facebook appartenant à des pages publiques et paramétrées comme publiques, les mêmes paramètres étant valables pour les commentaires. Cela permettra par la suite une redistribution libre du corpus. Cependant, afin d'assurer le plus de protection pour l'identité des utilisateurs – quoique les pseudonymes utilisés sur la plateforme ne correspondent pas forcément à leurs noms réels – nous avons décidé de désidentifier les commentaires. Ainsi, chaque nom d'utilisateur a été remplacé par un pseudonyme du type « L00001 ».

Cela nous a enfin permis de rassembler un corpus de 93 publications appartenant à trois pages Facebook différentes, avec un total de 20660 commentaires. Ce corpus se présente d'abord sous la forme de 93 fichiers CSV, qui permettent d'effectuer des modifications, ainsi que des recherches plus avancées à l'aide des scripts Python intégrant le module RegEx et les

librairies Panda. Cela pourrait servir, par exemple, à retrouver dans le corpus l'ensemble des variantes orthographiques de différents lexèmes. En outre, nous disposons du corpus dans une version qui se prête au traitement via le logiciel TXM, ce qui offre une meilleure visualisation des commentaires hiérarchisés, tout comme différentes possibilités d'annotation. Le corpus ainsi structuré se prêtera à une analyse autour de deux grands axes : le français de Côte d'Ivoire et le discours numérique. Il s'agira de déterminer quelles sont les stratégies que les locuteurs emploient afin de transposer à l'écrit un non-standard par excellence oral, mais aussi afin de mettre en évidence la variation régionale, qu'ils opposent ouvertement à une certaine image du français « neutre » ou « standard ». Ces stratégies, qui vont jusqu'à la représentation scripturale de la variation phonétique ou prosodique, jouent, évidemment, sur les ressources techniques inhérentes au discours numérique.

Bibliographie

Ahua, M. B. (2010). *Lexique illustré du nouchi ivoirien : quelle méthodologie ?*. Le français en Afrique, 25, 99-109.

Boutin, B. A., Kouadio, N. J. (2015). *Le nouchi c'est notre créole en quelque sorte, qui est parlé par presque toute la Côte d'Ivoire*. In Blumenthal, P. (éd.). *Dynamique des français africains : entre le culturel et le linguistique*. Peter Lang, 251-271.

Fairon, C., Klein, J.-R. (2010). *Les écritures et graphies inventives des SMS face aux graphies normées*. Le français aujourd'hui, 170, 113-122. 10.3917/lfa.170.0113.

Hammond, M. 2020. *Python for Linguists.* Cambridge University Press.

Millour, A. (2020). *Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées*. Thèse. Sorbonne Université.

Paveau, M-A. (2017). *L'analyse du discours numérique. Dictionnaire des formes et des pratiques*. Hermann.

La linguistique ergonomique au service de l'intelligibilité de la documentation prescriptive

MARTEL Eléna[1,2], CONDAMINES Anne[1], ARGUEL Amaël[1], KAHN Julien[2]

1 Laboratoire Cognition Langues Langage et Ergonomie (CLLE), 5 Allée Antonio Machado, 31100 Toulouse

{elena.martel, anne.condamines, amael.arguel}@univ-tlse2.fr

2 EDF R&D, 7 Boulevard Gaspard Monge, 91120 Palaiseau

{elena.martel, julien.kahn}@edf.fr

Mots-Clés : langue contrôlée, risque langagier, intelligibilité, documentation prescriptive

Dans le cadre d'une entreprise, il est essentiel que la documentation technique soit la plus intelligible possible pour les destinataires finaux. C'est le cas en particulier dans des contextes d'exploitation mettant en oeuvre des produits ou des procédés dangereux, afin de réduire les risques d'actions inadaptées. Dans ce contexte, nous étudions la conception d'une langue contrôlée. Les langues contrôlées se composent d'un ensemble de recommandations linguistiques destinées à renforcer la compréhension d'échanges oraux ou écrits (Kuhn, 2014). L'objectif est de diminuer le « risque langagier » i.e. un décalage qui s'installerait entre l'intention de communication du locuteur et la compréhension du message par le récepteur (Condamines, 2008). En effet, la clarté d'un texte professionnel est directement liée à la performance, évaluée selon qu'il déclenche ou non l'action qu'il était censé provoquer (Beaudet, 2001). L'objectif de cette thèse est d'établir des règles linguistiques à la fois performantes, efficaces et utilisables, basées sur l'analyse systématique d'un corpus de productions langagières réelles et sur l'évaluation de l'acceptabilité de ces règles par les utilisateurs.

Le corpus que nous étudions se caractérise par sa spécificité. En effet, il s'agit d'un document à caractère réglementaire de plusieurs chapitres et d'environ 300000 mots pour environ 6000 pages. Ce document formule l'ensemble des exigences techniques que les exploitants des installations doivent respecter afin de garantir la sûreté pour la production d'énergie. Ce document a été rédigé sans utilisation d'une langue contrôlée.

Nous cherchons dans un premier temps à identifier des motifs langagiers spécifiques au moyen du Traitement Automatique du Langage (TAL) pour déterminer les types d'informations transmises et leur mode de formulation, dans le but de formuler des alternatives langagières plus intelligibles et de les tester auprès des utilisateurs (Warnier, 2018). Longrée

et Mellet (2013) définissent les motifs comme étant « un cadre collocationnel » avec « un ensemble d'éléments fixes et de variables, susceptible d'accompagner la structuration textuelle, et simultanément, de caractériser des textes de genres divers ». Notre méthode repose sur l'analyse de corpus pour détecter des motifs récurrents. Pour extraire ces motifs, nous utiliserons des outils tel que *Sequential Data Mining under Constraints* (SDMC) (Béchet et al., 2013). Certains de ces motifs pourront apparaître comme entrainant potentiellement des difficultés de compréhension. Par exemple :

(1) « Si [condition], il est admis de réaliser des [opérations] sur le [système] ».

Dans le corpus, le motif « il est admis » est souvent utilisé pour indiquer une permission. Cependant, cette formulation peut entraîner des difficultés de compréhension puisqu'elle peut être interprétée comme une simple tolérance, plutôt qu'une permission explicite. Sur la base de nos connaissances linguistiques, nous proposerons des formulations alternatives dont l'acceptabilité sera évaluée auprès des utilisateurs. La technologie, en permettant l'extraction de motifs langagiers spécifiques via des outils comme le TAL, nous aide à créer des normes linguistiques pour la langue contrôlée. En ce sens, elle devient un véritable vecteur de productions langagière, car elle participe activement à la définition et à la standardisation de cette nouvelle langue contrôlée aux besoins du contexte professionnel.

Une des faiblesses des langues contrôlées existantes est que leur utilisabilité n'est que peu évaluée, c'est-à-dire que la compréhensibilité de la part des utilisateurs et donc l'efficacité dans la transmission de l'information ne sont que rarement prises en compte (Condamines, 2020). Nous sommes donc convaincus qu'impliquer les utilisateurs dans la conception et l'évaluation de la langue contrôlée présenterait un atout majeur pour garantir l'efficacité de la langue contrôlée proposée. Les utilisateurs seront ainsi plutôt des « acteurs de la conception » (Daniellou, 2004). Par exemple, les concepteurs pourront découvrir à la fois les contraintes des utilisateurs, tandis que les utilisateurs pourraient apprendre à partir du résultat des concepteurs (Béguin, 2004, 2008 ; Falzon, 2005). Il y a deux types de concepteurs au sein de notre étude, nous-mêmes qui sommes concepteurs de la langue contrôlée, et les rédacteurs techniques susceptibles d'utiliser la langue contrôlée pour rédiger la documentation. In fine, il s'agit d'articuler à la fois le point de vue des rédacteurs et celui des utilisateurs finaux (les opérateurs sur le terrain) pour concevoir cette langue contrôlée. Impliquer les rédacteurs permet de garantir la conservation du sens technique véhiculée à travers la documentation. Le fait d'impliquer les utilisateurs de la documentation permet d'en évaluer l'intelligibilité lors de sa mise en oeuvre en intégrant des méthodes de la psychologie cognitive et de l'ergonomie. Cette

évaluation peut se faire grâce à des expérimentations alliant linguistique et psychologie cognitive, par exemple via des indicateurs physiologiques ou de variables comportementales observés lors de mises en situations. L'acceptabilité par les rédacteurs et les opérateurs de la langue contrôlée sera faite en utilisant des outils d'évaluation de l'acceptabilité de type TAM (Davis, 1989).

Cette thèse en sciences du langage vise à mettre en oeuvre le TAL pour analyser et détecter des motifs langagiers dans un corpus de productions réelles, afin d'alimenter la réflexion linguistique dans la proposition de formulations alternatives. Les méthodes de la psychologie et de l'ergonomie cognitive seront aussi intégrées dans l'étude pour évaluer l'efficacité de ces recommandations linguistiques auprès des utilisateurs.)

Bibliographie

Beaudet, C. (2001). Clarté, lisibilité, intelligibilité des textes: un état de la question et une proposition pédagogique. *Recherches en rédaction professionnelle*, *1*(1), 1-19.

Béchet, N., Cellier, P., Charnois, T., Crémilleux, B., & Quiniou, S. (2013, January). SDMC: un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes. In *Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'13)*.

Béguin, P. (2004). L'ergonomie en conception. *Les nouveaux régimes de la conception, A. Hatchuel & B. Weill*.

Beguin, P. (2008). Conception et santé: quelques remarques sur le statut de l'activité de travail dans la conception des systèmes de production. *Psychologie du Travail et des Organisations*, *14*(4), 369-384. https://doi.org/10.1016/S1420-2530(16)30198-4

Condamines, A. (2008). Peut-on prévenir le risque langagier dans la communication écrite ?. *Langage & société*, (3), 77-97. https://doi.org/10.3917/ls.125.0077

Condamines, A. (2020). Towards an Ergonomic Linguistics. Application to the Design of Controlled Natural Languages. *International Journal of Applied Linguistics*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/full/10.1111/ijal.12313

Daniellou, F. (2004). L'ergonomie dans la conduite de projets de conception de systèmes de travail. *Ergonomie, 359-373.*

Davis, F. D. (1989). Technology acceptance model: TAM. *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption*, *205*, 219.

Falzon, P. (2005). Ergonomics, knowledge development and the design of enabling environments. In *Humanizing Work and Work Environment Conference* (pp. 10-12).

Kuhn, T. (2014). A survey and classification of controlled natural languages. *Computational linguistics*, *40*(1), 121-170. https://doi.org/10.1162/COLI_a_00168

Longrée, D., & Mellet, S. (2013). Le motif: une unité phraséologique englobante? Étendre le champ de la phraséologie de la langue au discours. *Langages*, (1), 65-79. https://doi.org/10.3917/lang.189.0065.

Warnier, M. (2018). Contribution de la linguistique de corpus à la constitution de langues contrôlées pour la rédaction technique. Doctoral Dissertation, University of Toulouse.

Un moteur de recherche sémantique : de l'extraction d'entités nommées à la création d'un RAG biographique

ROLIN Eva

CENTAL – UCLouvain – Louvain-la-Neuve (Belgique)

eva.rolin@uclouvain.be

Mots-Clés : Entités nommées, corpus, apprentissage actif, extraction d'informations, RAGs

La Biographie nationale est un dictionnaire biographique, publié entre 1866 et 1986 par l'Académie royale des sciences, des lettres et des beaux-arts de Belgique. Ce dictionnaire constitue une ressource essentielle pour les chercheurs, historiens, journalistes ou toute autre personne intéressée par l'histoire de la Belgique. Malheureusement, ces informations sont disponibles uniquement sous format papier, ce qui limite l'accès et rend la recherche d'informations plus difficile.

Pour faciliter l'accès à cette ressource, notre projet a pour objectif de développer un moteur de recherche biographique sous la forme d'un RAG (Retrieval Augmented Generation) augmenté d'informations biographiques structurées telles que les entités et les événements. D'un point de vue technique, notre recherche va donc s'intéresser principalement à l'extraction et la structuration des entités et des événements (dans un contexte de RAG). Toutefois, les ambitions applicatives de ce projet nous amènerons également à traiter des questions d'expérience utilisateur. En effet, les recherches menées en intelligence artificielle se concentrent bien souvent sur les seuls aspects techniques reléguant au second plan l'importance des utilisateurs finaux. Ces derniers doivent dès lors adapter la technologie à leurs contextes spécifiques. Nous pensons toutefois qu'il est primordial de remettre l'‘utilisateur au centre du processus de conception pour comprendre ses besoins et les concilier avec les contraintes techniques, garantissant ainsi des solutions efficaces et dignes de confiance.

À ce stade du projet, notre recherche s'est principalement concentrée sur la reconnaissance des entités nommées et l'établissement d'un corpus d'entrainement.

Bien que les entités nommées soient un sujet de recherche bien établi et largement étudié, comme en attestent plusieurs états de l'art (Nadeau & Sekine, 2007 ; Yadav & Bethard, 2019 ; Li et al., 2020 ; Keraghel et al., 2024), les travaux spécifiquement centrés sur le français restent relativement limités (Béchet & Charton, 2010 ; Stern & Sagot, 2010 ; Maurel et al., 2011, Nouvel & Soulet, 2011 ; Dupont & Tellier, 2014 ; Dupont, 2017). Un défi majeur concerne la

reconnaissance des entités nommées complexes, telles que les entités imbriquées (par exemple, [conservateur du [Musée [Plantin pers] org] prof]) et discontinues (par exemple, [ministre [des finances coord] et [de l'économie coord] prof]). Pour aborder ces structures complexes, à l'instar de Finkel et Manning (2009), nous proposons de substituer à la tâche standard de segmentation, une tâche d'analyse syntaxique en dépendances (pour les entités uniquement). En effet, actuellement, le formalisme standard d'annotation (IOB2, pour *Inside-Outside-Begin*) se limite à la délimitation des frontières d'entités et ne rend pas compte de leur structure syntaxique interne. Or, cette structure est essentielle au traitement des entités imbriquées et discontinues.

La complexité du formalisme étudié rend le travail d'annotation des données d'entrainement particulièrement fastidieux. Pour accélérer ce processus, nous avons recours à l'apprentissage actif (Settles, 2009), qui permet d'optimiser la sélection d'exemples, réduisant ainsi la quantité de données à annoter. Peu de recherches ont exploré cette approche appliquée aux entités nommées en français (Claveau & Kijak, 2015 ; Naguib et al., 2023). Afin de déterminer la stratégie de sélection et les mesures d'agrégation les plus efficaces, nous avons simulé une boucle d'apprentissage. Nous avons ainsi pu observer que la méthode par moindre confiance semble obtenir les meilleures performances. A ce stade, le processus d'annotation est toujours en cours. Toutefois, nos premières analyses montrent déjà un accord inter-annotateurs de 71.5 %.

Loin d'être arrivée à son terme, notre recherche doit encore traiter deux questions essentielles : comment valoriser les informations extraites au sein de RAGs et quelle place donner à l'utilisateur.

Récemment, certains chercheurs ont suggéré de conférer aux modèles la capacité d'accéder à la mémoire externe afin qu'ils puissent obtenir davantage d'informations dans le processus de génération (Li et al., 2022). Plutôt que de se limiter à ce qu'ils ont appris, les modèles peuvent récupérer des informations à partir de sources supplémentaires telles que des corpus d'entrainement, des données externes ou des données non-supervisées (Li et al., 2022). Ces données peuvent provenir de diverses sources et types. Par exemple, des bases de données spécialisées comme *PubMed* peuvent fournir des informations médicales actualisées, ce qui peut améliorer l'exactitude des contenus destinés aux professionnels de la santé (Thomo, 2024). Les plateformes d'actualités telles que *Bloomberg* et *Reuters* et les médias sociaux comme *Twitter* et *Reddit* peuvent enrichir l'analyse des sentiments dans le domaine financier (Zhang et al., 2023). Les graphes de connaissances (Jiang et al., 2023) et les moteurs de recherche comme *Bing* et *Google* permettent d'accéder à des informations en temps réel, ce qui peut

également être bénéfique pour divers domaines de recherche (Lazaridou et al., 2022). Il s'agira donc pour nous de structurer les données extraites de manière à alimenter le modèle et guider la génération.

L'utilisateur doit être au centre de la conception de notre moteur de recherche sémantique. Cette approche est essentielle pour garantir que l'outil réponde de manière adéquate aux besoins des utilisateurs, qui sont les principaux concernés. Il est donc important que la conception et l'utilisation de l'outil adoptent une approche empirique, basée sur l'expérience des utilisateurs, plutôt qu'une approche rationaliste centrée uniquement sur les exigences de performance (Shneiderman, 2022). Le Human Centered Design (HCD) et la Human Centered AI (HCAI), une extension du HCD spécifiquement appliquée aux systèmes d'intelligence artificielle, peuvent offrir des solutions efficaces à ces préoccupations. Le HCD est une approche de conception qui place les besoins, les comportements et les attentes des utilisateurs finaux au coeur du processus de conception. Il implique plusieurs étapes clés : l'engagement actif des utilisateurs pour comprendre leurs besoins et les tâches à accomplir, la génération d'idées pour développer des solutions possibles, la création de prototypes pour tester ces solutions auprès des utilisateurs, et l'itération en fonction de leurs retours jusqu'à ce que le produit final réponde pleinement à leurs attentes (Maguire, 2001). Cette approche, associée à des techniques de collecte et d'analyse UX (User eXperience), guidera nos choix méthodologiques.

Bibliographie

Béchet, F., & Charton, E. (2010, March). Unsupervised knowledge acquisition for extracting named entities from speech. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 5338-5341). IEEE. https://doi.org/10.1109/ICASSP.2010.5494962

Claveau, V., & Kijak, E. (2015, June). Stratégies de sélection des exemples pour l'apprentissage actif avec des champs aléatoires conditionnels. In Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs (pp. 13-24).

Dupont, Y., & Tellier, I. (2014, July). A Named Entity recognizer for French (Un reconnaisseur d'entités nommées du Français) [in French]. In Proceedings of TALN 2014 (Volume 3: System Demonstrations) (pp. 40-41).

Dupont, Y. (2017). La structuration dans les entités nommées (Doctoral dissertation, Université Sorbonne Paris Cité).

Finkel, J. R., & Manning, C. D. (2009, August). Nested named entity recognition. In Proceedings of the 2009 conference on empirical methods in natural language processing (pp. 141-150).

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., ... & Neubig, G. (2023). Active retrieval augmented generation. arXiv preprint arXiv:2305.06983. https://doi.org/10.48550/arXiv.2305.06983

Keraghel, I., Morbieu, S., & Nadif, M. (2024). A survey on recent advances in named entity recognition. arXiv preprint arXiv:2401.10825. https://doi.org/10.48550/arXiv.2401.10825

Lazaridou, A., Gribovskaya, E., Stokowiec, W., & Grigorev, N. (2022). Internet-augmented language models through few-shot prompting for open-domain question answering. arXiv preprint arXiv:2203.05115. https://doi.org/10.48550/arXiv.2203.05115

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. IEEE transactions on knowledge and data engineering, 34(1), 50-70. http://dx.doi.org/10.1109/TKDE.2020.2981314

Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A survey on retrieval-augmented text generation. arXiv preprint arXiv:2202.01110. https://doi.org/10.48550/arXiv.2202.01110

Maguire, M. (2001). Methods to support human-centred design. International journal of human-computer studies, 55(4), 587-634. https://doi.org/10.1006/ijhc.2001.0503

Maurel, D., Friburger, N., Antoine, J. Y., Eshkol, I., & Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. Revue TAL: traitement automatique des langues, 52(1), 69-96.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26. https://doi.org/10.1075/li.30.1.03nad

Naguib, M., Névéol, A., & Tannier, X. (2023). Stratégies d'apprentissage actif pour la reconnaissance d'entités nommées en français. In 18e Conférence en Recherche d'Information et Applications--16e Rencontres Jeunes Chercheurs en RI--30e Conférence sur le Traitement Automatique des Langues Naturelles--25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (pp. 232-247). ATALA.

Nouvel, D., & Soulet, A. (2011, January). Annotation d'entités nommées par extraction de règles de transduction. In Extraction et la Gestion des Connaissances (p. 119).

Settles, B. (2009). Active learning literature survey.

Shneiderman, B. (2022). Human-centered AI. Oxford University Press.

Stern, R., & Sagot, B. (2010, July). Détection et résolution d'entités nommées dans des dépêches d'agence. In Traitement automatique des langues naturelles: Taln 2010. https://doi.org/10.1051/cmlf/2010217

Thomo, A. (2024). PubMed Retrieval with RAG Techniques. Studies in health technology and informatics, 316, 652-653. https://doi.org/10.3233/shti240498

Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. arXiv preprint arXiv:1910.11470. https://doi.org/10.48550/arXiv.1910.11470

Zhang, B., Yang, H., Zhou, T., Ali Babar, M., & Liu, X. Y. (2023, November). Enhancing financial sentiment analysis via retrieval augmented large language models. In Proceedings of the fourth ACM international conference on AI in finance (pp. 349-356). https://doi.org/10.48550/arXiv.2310.04027

Using automatic annotation tools to analyze inter- and intra- speaker variation in Australian English

MAS Erwanne

Laboratoire CLLE, UMR 5263 CNRS & U. Toulouse Jean Jaurès

erwanne.mas@univ-tlse2.fr

Mots-Clés : Automatic annotation, Sociophonetics, Australian English

This study evaluates the performance of automatic annotation tools, specifically the Montreal Forced Aligner, to analyze inter- and intra- speaker variation in spoken Australian English. Forced alignment, which is a computational process used in phonetics to automatically match spoken audio with its corresponding transcription at the word and phoneme levels, is widely used in sociophonetic research to segment speech into phoneme-level units, reducing the time and effort required for manual annotation (Gut & Voormann, 2014). However, the effectiveness of forced annotation tools varies depending on factors such as speech type, audio quality, and speaker characteristics. This study focuses on a corpus of Australian speech recorded between 2003 and 2023 across Queensland, New South Wales, and Victoria (Mas & Przewozny, 2023). The Montreal Forced Aligner was used to automatically time-align spontaneous speech and reading tasks, followed by manual verification to assess transcription accuracy (Mcauliffe et al., 2017).

Our findings show that the aligner performed well on reading tasks, but struggled with spontaneous speech, particularly in segments involving multiple speakers and fast speech rates. On average, 50% of the alignments required manual corrections, highlighting the challenges posed by Australian English's unique phonetic features, such as regional vowel differences (Cox & Palethorpe, 2019). Furthermore, the study examines how both speaker diarization, i.e., the process of identifying and distinguishing between different speakers in an audio recording and overall speech rate, influence the performance of the aligner for automatic annotations of underrepresented English varieties such as Australian English (Mackenzie & Turton, 2020).

By comparing segments involving multiple and individual speakers, as well as speakers of different ages and genders, we demonstrate how forced alignment tools can be handled to better capture speaker variation and phonetic diversity. These insights can help linguists develop more efficient methods to process spoken data, particularly in sociophonetic studies of Australian English (Cox & Docherty, 2024; Gonzalez et al., 2020).)

Bibliographie

Cox, Felicity, Palethorpe, Sallyanne, 2019, “Vowel variation across four major Australian cities”, dans Calhoun, S., Escudero, P., Tabain, M. et Warren, P. (éds.) *Proceedings of the International Congress of Phonetic Sciences 2019*, Canberra, Australia: Australasian Speech Science and Technology Association Inc. p. 577-581.

Cox, Felicity, Docherty, Gerard, 2024, “Sociophonetics and vowels”, dans Strelluf, Christopher. (éd) *Routledge Handbook of Sociophonetics*, Abingdon, Routledge, p. 114-142.

Gonzalez, Simon, James Grama, & Travis, Catherine 2020. « Comparing the performance of forced aligners used in sociophonetic research ». *Linguistics Vanguard* 6(1), 20190058. https://doi.org/10.1515/lingvan-2019-0058

Gut Ulrike et Voormann Holger. (2014). « Corpus Design », in J. Durand, U. Gut et G. Kristoffersen (éd.). *The Oxford Handbook of Corpus Phonology*. Oxford : Oxford University Press, 13-26.

Mackenzie, Laurel, Turton, Danielle. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard*, *6*(s1), Article 20180061. *https://doi.org/10.1515/lingvan-2018-0061*

Mas, Erwanne, Przewozny, Anne, 2023, Corpus *Phonology of Contemporary English: Australia*. CLLE (CNRS UMR 5263) ; LPL (CNRS UMR 7309) ; CREA (EA 370) ; CLILLAC-ARP (EA 3967) https://doi.org/10.34847/COCOON.AB9E1776-BD3F-44E4-9E17-76BD3FB4E4BE

Mcauliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, Morgan Sonderegger, 2017. « Montreal Forced Aligner: Trainable text-speech alignment using Kaldi ». *Proceedings of the 18th Conference of the International Speech Communication Association.*