Laboratoire
Linguistique, Langues, Parole
LILPA | UR 1339
Université de Strasbourg

Corpus et linguistique outillée

Séminaire inter-axes 4 avril 2025

Amphi 23 Escarpe

Organisé par Monika Pukli (Axe 2), Julia Putsche (Axe 3) & Hélène Vassiliadou (Axe 1)

Programme

15h00-15h20 : *Corpus et linguistique outillée : présentation*
Monika Pukli, Julia Putsche & Hélène Vassiliadou

15h20-15h50 : *Des discours institutionnels aux pratiques de classe : de l'utilité d'un corpus hétérogène*, Lucile Hamm (doctorante, axe 3)

15h50-16h20 : *Aborder l'énonciation sur corpus oral : approches et outils*, Igor Ilic (doctorant, axe 2)

16h20-16h50 : *GREMO-LING : un corpus, plusieurs axes d'analyse*, Thalassio Briand, Salomé Klein, Joé Laroche, Louis Raguenet (doctorants, axe 1)

PAUSE : 16h50-17h00

17h00-17h30 : *La création des corpus pédagogiques thématiques pour l'enseignement des langues*, Szilvia Szita (enseignante contractuelle, Dr. en Didactique des langues, GEO, associée à l'Axe 3)

17h30-18h00 : *La linguistique outillée au service de l'analyse littéraire : la manducation chez Zola*, Francine Gerhard-Krait (MCF) et Marie Lammert (PR) (Axe 1)

Résumés des communications

Corpus et linguistique outillée : présentation

Monika Pukli, Julia Putsche & Hélène Vassiliadou

La disponibilité de données langagières (surtout numérisées), produites dans différents contextes de communication, a modifié les pratiques de nombreux linguistes. Ces collections de textes, appelées corpus, rendent possible l'observation à grande échelle d'usages langagiers diversifiés. Le terme de *corpus* renvoie ainsi à une grande collection de textes authentiques censée être représentative d'une langue ou d'une variété de langues. Des outils permettent de les explorer et d'en extraire des informations variées. L'objectif de ce séminaire Inter-axes est de permettre aux étudiants d'observer les méthodes d'exploration et d'analyse des corpus et de voir comment nous pouvons les mettre en œuvre en utilisant, en fonction de la question linguistique posée, différents outils. Nous souhaitons ainsi exposer en quoi consiste une démarche de recherche fondée sur l'analyse outillée de ces données, et à réfléchir sur la nature des données elles-mêmes (format, annotation, documentation des corpus).

Références indicatives

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2), 141-161.

Borillo, A. (1996), Exploration automatisée de textes de spécialité : repérage et identification de la relation lexicale d'hyperonymie, *LINX* 34-35, 113-124.

Boulton, A. & Tyne, H. (2014). *Des Documents authentiques aux corpus : démarches pour l'apprentissage des langues*. Paris : Didier.

Cislaru, G. et Sitri, F. (2009), TEXTE ET DISCOURS. Corpus, co-texte et analyse automatique du point de vue de l'analyse de discours, *Corpus* 8, 85-104.

Fillmore, C. (1992), Corpus Linguistics or Computer-aided armchair linguistics, *Directions in Corpus Linguistics*, Berlin, Mouton de Gruyter, 35-60.

Gries, S. T., & Newman, J. (2014). Creating and using corpora. *Research methods in Linguistics*, 257-287.

Habert, B., Nazarenko, A., Salem, A. (1997), *Les linguistiques de corpus*, Paris, Armand Colin.

Hearst, M. (1992), Automatic Acquisition of Hyponyms from Large Texte Corpora, *COLING* 92, 539-545.

Jacques, M.P. (2016), Une linguistique outillée, pour quels objets ?, *Histoire, épistémologie, Langage* 38(2), 87-99.

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1), 2001, pp. 97-133

Lebart, L., Pincemin, B. et Poudat, C. (2019) *Analyse des données textuelles*. Presses Universitaires du Québec.

Lecolle, M. (2007), Polysignifiance du toponyme, historicité du sens et interprétation en corpus. Le cas de Outreau, *Corpus* 6, 101-125.

Leech, G. (1992), Corpora and theories of linguistic performance, *Directions in Corpus Linguistics*, Berlin, Mouton de Gruyter, 105-122.

Léon, J. (2008), Aux sources de la Corpus Linguistics : Firth et la London School, *Langages* 171, 12-33.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Pêcheux, M. (1984), Sur les contextes épistémologiques de l'analyse de discours, *Mots* 9, 7-17.

Pincemin, B. (2007), Concordances et concordanciers : de l'art du bon KWAC, *XXVIIe colloque d'Albi*, 33-42.

Poudat C., Landragin, F. (2017). *Explorer un corpus textuel : Méthodes, pratiques, outils*. DeBoeck, Champs linguistiques.

Teubert, W. (2005), My version of corpus linguistics, *International Journal of Corpus Linguistics* 10(1), 1-13.

Des discours institutionnels aux pratiques de classe : de l'utilité d'un corpus hétérogène

Lucile Hamm

À la croisée entre la didactique des langues, la sociolinguistique et la dialectologie, le projet de thèse de doctorat intitulé « L'alsacien à l'école ? Questions didactiques et de politiques linguistiques » vise à interroger l'articulation entre le discours institutionnel autour de la « langue régionale » (allemand standard et alsacien) portant sur le domaine éducatif/scolaire en Alsace et les pratiques d'enseignement de l'alsacien/en alsacien, en particulier à l'école préélémentaire.

Afin de constituer notre corpus, nous avons procédé au recueil de données et d'observables dans différents fonds d'archives institutionnels et dans plusieurs contextes gravitant autour du parcours allemand-alsacien-français « Tomi Ungerer » proposé à titre expérimental depuis septembre 2023 dans des classes de maternelle de l'académie de Strasbourg.

Nous aborderons la notion d'« hétérogénéité » (Krieg-Planque, 2012, p. 115), qui nous a guidée dans le processus de constitution du corpus. En effet, la multiplicité des espaces et contextes de collecte d'observables, ainsi que la démarche ethnographique adoptée (Brougère et al., 2015, p. 10) se retranscrivent dans la nature des éléments retenus pour l'analyse : procès-verbaux de délibérations, comptes rendus de réunions, notes de service, rapports institutionnels, transcriptions de formations, grilles d'observation participante de classe, réponses à des questionnaires et prises de vues.

En guise d'ouverture, nous évoquerons la manière dont nous envisageons d'articuler l'analyse de ces éléments, en lien avec nos objectifs et questions de recherche.

Bibliographie

Brougère, G., Kubanek, A., Macaire, D., & Putsche, J. (2015). *La valisette franco-allemande : quelle place pour la langue et la culture de l'autre à l'école maternelle et au « Kindergarten » ?* Office franco-allemand pour la Jeunesse.

Krieg-Planque, A. (2012). *Analyser les discours institutionnels*. Armand Colin.

Aborder l'énonciation sur corpus oral : approches et outils

Igor ILIC

L'existence du *discours oral* en tant que donnée observable est discutée, entre autres, par Kuyumcuyan (2002) et Simon (2004). L'une met la focale sur le dialogue en tant que pratique langagière *vs* en tant que texte achevé, l'autre montre qu'un mode d'accès particulier, l'*expérienciation*, permet à l'analyste de saisir les deux niveaux évoqués. Cette discussion ouvre, au moins, deux voies. La première concernerait le débat sur l'immanence de l'objet observé et, dans ce contexte, celui entre une linguistique textuelle et une analyse du discours. Nous ne prendrons pas cette voie. Le lecteur curieux se reportera, pour des synthèses du sujet, à Auchlin *et al.* (2004) ou Roulet *et al.* (2001). La seconde voie concernerait les répercussions de ces visions sur le traitement des données issues des interactions verbales. C'est la voie que nous explorerons dans cette brève

communication à travers trois points. La prise en compte des particularités de l'*interaction* comme forme d'organisation du discours pour un phénomène énonciatif tel que le discours représenté. L'élaboration d'un guide d'annotation en discours représenté permettant à l'annotateur d'annoter à partir de la transcription et de l'enregistrement audio d'une interaction. L'utilisation de l'outil d'annotation INCEpTION et les défis de la variation des phénomènes adjacents comme les particules ou les décalages énonciatifs. *In fine*, des pistes quant à l'exploitation et l'extraction des données annotées pourront être évoquées.

Bibliographie

Auchlin, A., Filliettaz, L., Grobet, A., & Simon, A.-C. (2004). (En)action, expérienciation du discours et prosodie. *Cahiers de linguistique française*, *26*, 217-249.

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION Platform : Machine-Assisted and Knowledge-Oriented Interactive Annotation. In D. Zhao (Éd.), *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations* (p. 5-9). Association for Computational Linguistics.

Kuyumcuyan, A. (2002). *Diction et mention : Pour une pragmatique du discours narratif*. P. Lang.

Roulet, E., Filliettaz, L., & Grobet, A. (2001). *Un modèle et un instrument d'analyse de l'organisation du discours*. Lang.

Simon, A. C. (2004). *La structuration prosodique du discours en français : Une approche multidimensionnelle et expérientielle*. P. Lang.

GREMO-LING : un corpus, plusieurs axes d'analyse

Thalassio Briand, Salomé Klein, Joé Laroche, Louis Raguenet

Le corpus GREMO-LING regroupe des entretiens de patients atteints de lésion cérébrale acquise (LCA) menés à différentes étapes de leur accompagnement socio-médical et thérapeutique. L'objectif de nos recherches est de proposer des critères d'évaluation linguistiques de la thérapie de régulation émotionnelle GREMO suivie par les patients (basée sur la TCD, Linehan 2017). Le corpus GREMO-LING présente plusieurs défis, parmi lesquels le traitement d'un grand corpus oral, sa compatibilité avec les différentes méthodologies adoptées et la garantie du secret médical. L'enjeu est dès lors de définir les besoins de chaque sous-discipline linguistique (TAL, organisation de l'oral, agentivité, implicite…) et d'adapter le corpus pour qu'il devienne une ressource exploitable par l'ensemble de l'équipe de recherche.

Références

BATTISTELLI, D., ÉTIENNE, A., LECORVÉ, G. (2022), L'émotion à un niveau textuel : La fonction structurante des émotions observée à partir d'annotations, *Discours. Revue de linguistique, psycholinguistique et informatique*, *30*.

BAYERL, P. S., PAUL, K. I. (2011), What Determines Inter-Coder Agreement in Manual Annotations ? A Meta-Analytic Investigation, *Computational Linguistics*, *37*(4), 699–725.

CRIBLE, L., DIDIRKOVÁ, I., DODANE, C., KOSMALA, L. (2022). Towards an inclusive system for the annotation of (dis)fluency in typical and atypical speech, *Clinical Linguistics & Phonetics*, *38*(*4*), 1–18.

ÉTIENNE, A., BATTISTELLI, D. (2021), *Annotation manuelle des émotions dans des textes écrits avec la plateforme Glozz.* [Research Report], MoDyCo ; Paris : Université Paris Nanterre.

BRIAND, T., FAUTH, C., KUPPELIN, M. (2024), Intonation and Fluency in Emotionally Dysregulated French Patients with an Acquired Brain Injury: Case Studies, *in Proc. Speech Prosody 2024*, 807–811.

BRIAND, T., FAUTH, C., VASSILIADOU, H. (2022), Marques de l'émotion dans la fluence d'un patient cérébrolésé : Étude préliminaire de faisabilité, *in XXXIVe Journées d'Études Sur La Parole – JEP 2022*, International Speech Communication Association, 90–98.

KLEIN, S., BENNINGER, C., BRIAND, T., FAUTH, C., GERHARD-KRAIT, F., KUPPELIN, M., LAMMERT, M., LAROCHE, J., RAGUENET, L., SCHNEDECKER, C., TODIRASCU, A., KRASNY-PACINI, A., VASSILIADOU, H. (à paraître), *Projet GREMO-LING. Annotation manuelle de l'expression de l'émotion dans des transcriptions de l'oral. Guide d'annotation pour les transcriptions GREMO*.

KLEIN, S., TODIRASCU, A., VASSILIADOU, H., KUPPELIN, M., BECART, J., BRIAND, T., CORIDON, C., GERHARD-KRAIT, F., LAROCHE, J., ULRICH, J., KRASNY-PACINI, A. (2024), Annotating Emotions in Acquired Brain Injury Patients' Narratives, *in* D. Demner-Fushman, S. Ananiadou, P. Thompson, & B. Ondov (eds), *Proceedings of the First Workshop on Patient-Oriented Language Processing* (*CL4Health*) *@ LREC-COLING 2024, ELRA and ICCL*, , ACL Anthology, 26–36 (aclanthology.org/2024.cl4health-1.4.pdf).

LINEHAN, M. M., KORSLUND, K. E., HARNED, M. S., GALLOP, R. J., LUNGU, A., NEACSIU, A. D., MCDAVID, J., COMTOIS, K. A., MURRAY-GREGORY, A. M. (2015), Dialectical Behavior Therapy for High Suicide Risk in Individuals With Borderline Personality Disorder: A Randomized Clinical Trial and Component Analysis, *JAMA Psychiatry*, *72*(5), 475–482.

La création des corpus pédagogiques thématiques pour l'enseignement des langues

Szilvia Szita

Un des objectifs des niveaux élémentaire et indépendant du Cadre européen commun de référence pour les langues (CECRL) est d'amener l'apprenant — alors qu'il n'est pas encore un utilisateur autonome de la langue — à être capable d'interagir dans les situations authentiques de la vie quotidienne. Or, la plupart des matériaux pédagogiques consacrés à ces niveaux ne présentent qu'un échantillonnage partiel et souvent incomplet du langage authentique. Dans cette présentation, nous proposons une méthodologie dédiée, d'une part, à la construction de corpus thématiques à des fins pédagogiques aux niveaux A1+ à B1 du CECRL et, d'autre part, à l'identification du vocabulaire-clé associé.

Notre approche repose sur une sélection thématique de textes allemands selon les sujets définis par le CECRL afin de créer (1) un corpus ciblé et (2) de taille relativement faible, (3) regroupant des textes (quasi-)authentiques, adaptés aux niveaux des apprenants. En accord avec les résultats issus de la linguistique de corpus, soulignant la place centrale des unités multi-lexicales (UML) dans la construction du vocabulaire, le vocabulaire-clé tiré de cette collection consiste en UML plutôt qu'en mots individuels. Le procédé de sélection est illustré à travers l'exemple concret d'un sous-thème de la catégorie « Essen und Trinken » (Nourriture) du CECRL formé à partir d'un corpus de critiques de restaurants. Nous concluons par le constat que cette approche empirique peut mener à un changement dans la perception de la nature du vocabulaire-clé susceptible d'être présenté aux apprenants aux niveaux de compétences linguistiques inférieurs.

Références

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, 17, 47–64.

Cavalla, C. (2019). Corpus numériques : critères pour l'enseignement des langues. Entre présence et distance. Enseigner et apprendre les langues à l'université à l'ère numérique, Equipe PERL, Paris.

Chambers, A. (2019). Towards the corpus revolution? Bridging the research-practice gap. *Language Teaching*, 52(4), 460–475.

Cobb, T. & Boulton, A. (2015). Classroom applications of corpus analysis. In Biber, D. & Reppen, R. (dir.), *The Cambridge handbook of English corpus linguistics*. Cambridge University Press.

De Fornel, M. & Verdier, M. (2018). Corpus, classes d'exemples et collections en analyse de conversation. *Corpus*, 18, 1–15.

Forti, L. & Spina, S. (2019). Corpora for linguists vs. corpora for learners: Bridging the gap in Italian L2 learning and teaching. *EL.LE*, 8(2), 349–362.

Reppen, R. (2016). Designing and building corpora for language learning. In Farr, F. & Murray, L. (éds). *The Routledge handbook of language learning and technology*. Routledge.

Schaeffer-Lacroix, E. (2012). Qu'est-ce qui rend les corpus 'pédagogiques' ? *Procedia - Social and Behavioral Sciences* 34, 198–201.

Scharloth, J. ; Okamura S. ; Lange, W. (2016). *Gibt es einen Kernwortschatz ? Datengeleitete Perspektiven auf die Erstellung von Grundwortschätzen für Deutsch als Fremdsprache*. Rapport du projet de recherche Grant Aid for Scientific Research, Japanese Society for the Promotion of Science 2011–2015.

La linguistique outillée au service de l'analyse littéraire : la manducation chez Zola

Francine Gerhard-Krait & Marie Lammert

Chez Zola, la manducation est un marqueur du déterminisme social des personnages. L'utilisation d'un logiciel de textométrie comme TXM permet d'identifier et d'annoter les verbes de manducation tout en les mettant en relation avec les différents personnages présents dans les Rougon-Macquart. Nous mettrons en évidence les ressorts méthodologiques d'une telle investigation qui nécessite une analyse fine des verbes de manducation en termes de traits sémantiques permettant d'opérer des croisements et d'élaborer des types de manducation. L'outil permettra également de confronter les différents volumes, ainsi que certains personnages.

Références

Fellbaum, C. (Éd.). (1998). *WordNet, An Electronic Lexical Database* (The MIT Press). Cambridge, Massachusetts; London, England.

Gerhard-Krait, F. & Lammert, M. (2020). Identités sociale et individuelle dans Les Rougon-Macquart : étude outillée des verbes de manducation, *in* B. Marquer (éd.), « Dis-moi ce que tu manges, je te dirai ce que tu es ». *Fictions identitaires, fictions alimentaires*, Strasbourg, Presses Universitaires de Strasbourg, 83-103.

Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement (Vol. 2, p. 1021-1032). Présenté à 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010, Edizioni Universitarie di Lettere Economia Diritto. Consulté à l'adresse https://halshs.archives-ouvertes.fr/halshs-00549779/document

Lair, A. V. (2003). *Les arts de la table : Nourriture et classes sociales dans la litterature Francaise du dix-neuvième siècle*. The Ohio State University.

Sánchez Cárdenas, B. (2011). Structuration Hiérarchique du lexique verbal à travers la propriété de troponymie. *Revista de Lingüística y Lenguas Aplicadas*, 6(1), 329-340. https://doi.org/10.4995/rlyla.2011.913

Sicotte, G. (2003). Une petite histoire du motif du repas au XIX^e siècle. *Textyles. Revue des lettres belges de langue française*, (23), 10-19.