Colloque

ConCorDiaL 2024

Constitution de Corpus en Diachronie Longue

Entre tradition philologique et analyse quantitative

7 - 8 novembre 2024

ENS de Lyon
Bât. Buisson
Salle D8-001

ihrim · icar · LIDILEM Université Grenoble Alpes · CERCC · ENS DE LYON

PROGRAMME

Jeudi 7 novembre

9h30 – 10h00 Accueil—Petit déjeuner

10h00 – 11h00 / Conférence 1

Au(x) fil(s) du temps - Analyse de données diachroniques en linguistique de corpus et textométrie, Sascha Diwersy (U. de Montpellier, PRAXILING)

11h00 – 11h35 Eszter Kovács (Vrije Universiteit Brussel)

L'ancienne sagesse et la naissance des sciences modernes : constitution des corpus à plusieurs niveaux

11h35 – 11h50 Pause café

11h50 – 11h25 David Kahn (INU Champollion, FRAMESPA), Sascha Diwersy (U. de Montpellier 3, PRAXILING)

L'évolution des mots de l'hétérodoxie dans l'Espagne du XVI^{e} siècle. Éditorialisation outillée et exploration textométrique d'une collection échantillonnée de sources inquisitoriales espagnoles et d'ouvrages de spiritualité

12h25 – 13h00 Tanguy Lemoine (U. de Grenoble Alpes, Litt&Arts)

Étudier la variation générique en diachronie : le point-virgule comme indice linguistique de différenciation des genres littéraires aux XVI^{e} et XVII^{e} siècles

13h00 – 14h15 déjeuner

14h15 – 14h50 Denisa Bumba, Gabriella Parussa (Sorbonne Université, STIH)

Constitution d'un corpus longitudinal par la transcription automatique (HTR/OCR) : problèmes liés à l'héterogénéité des sources

14h50 – 15h25 Alice Brenon (INSA Lyon, LIRIS & U. de Lyon 2, ICAR)

D'une encyclopédie à l'autre: le corpus «parallèle» du projet GEODE

15h25 – 16h00 Lucas Lévêque, Florian Cuny, Noé Gasparini (chercheurs indépendants)

La Dicothèque : un outil pour l'exploration pluridimensionnelle des dictionnaires

16h00 – 16h30 Pause café

16h30 – 17h30 / Conférence 2

L'IA peut-elle vraiment nous aider à explorer les grands corpus littéraires ?
Thierry Poibeau (CNRS, LATTICE)

Cocktail

Vendredi 8 novembre

9h30 – 10h30 / Conférence 3

Le projet Open French Corpus : constituer un corpus à partir de l'existant,
Céline Poudat (U. de Nice, BCL)

10h30 – 11h05 Lucence Ing (École nationale des chartes, Centre Jean Mabillon)

La constitution d'un corpus pour l'analyse des disparitions lexicales (XIII^{e}-XV^{e} siècles) : approches computationnelles et qualitatives

11h05 – 11h30 Pause café

11h30 – 12h05 Mathieu Dehouck, Sophie Prévost (CNRS, LATTICE), Mathilde Regnault (U. de Stuttgart), Loïc Grobol (U. Paris Nanterre, MODICO)

Comparaison de deux approches pour l'analyse syntaxique du français et du latin en diachronie

12h05 – 12h40 Natasha Romanova, Rayan Ziane (U. de Caen Normandie, CRISCO)

Quelques pistes pour surmonter les contraintes pour l'annotation syntaxique de corpus en diachronie longue

12h40 – 13h00 Discussion générale et clôture

13h00 – 14h00 Déjeuner

Résumés

Jeudi 7 novembre, 10h00

Au(x) fil(s) du temps - Analyse de données diachroniques en linguistique de corpus et textométrie

Sascha Diwersy (PRAXILING, Université Paul Valéry - Montpellier 3)

L'objectif de notre communication est de donner une synthèse des principales méthodes utilisées en linguistique de corpus et en textométrie pour l'exploration de bases textuelles structurées sous forme de séries chronologiques. Notre illustration passera par la mise en évidence d'un ensemble de résultats issus de différentes études s'inscrivant dans le domaines de la morphosyntaxe (Bres et al., 2018), de la sémantique lexicale (Diwersy et al., 2021 ; Diwersy & Verine, 2021) et de l'Analyse du discours (Diwersy & Luxardo, 2016 ou encore Diwersy et al., 2020). L'exposé se conclura par la présentation d'outils mettant en oeuvre différents procédés d'exploration diachronique comme le *Variability -based Neighbor Clustering* (VNC, Gries & Hilpert, 2008 ; Hilpert & Gries, 2009) ou les calculs de spécificités connexes (Salem 1988, 126-132) et d'accroissements spécifiques (Lebart & Salem 1994, 221-224).

Références

BRES Jacques, DIWERSY Sascha et LUXARDO Giancarlo (2018) « The competition between present conditional and prospective imperfect in French over the centuries », In Ayoun, D., Celle, A. & Lansari, L. (eds.), *Tense, Aspect, Modality and Evidentiality: Cross-linguistic perspectives*. Amsterdam: John Benjamins, 65-80.

DIWERSY Sascha, JACKIEWICZ Agata, LUXARDO Giancarlo et STEUCKARDT Agnès (2021) « Les sens de “numérique” : émergence d’emplois et dynamique du changement sémantique », *Linx*, 82. ‹DOI : 10.4000/linx.8153›.

DIWERSY Sascha, JAY ROBERT Pierre LEANDRO Camila, STEUCKARDT Agnès et CHANDELIER Marie (2020) Entre contrôle et protection – L’évolution de la représentation des insectes dans le discours médiatique en France. In *JADT’2020 : Actes des 15èmes Journées Internationales d’Analyse statistique des Données Textuelles*, Toulouse, France.

DIWERSY Sascha et LUXARDO Giancarlo (2016) « Mettre en évidence le temps lexical dans un corpus de grandes dimensions : l’exemple des débats du Parlement européen », In *JADT 2016 : Actes des 13èmes Journées internationales d’Analyse statistique des Données Textuelles*, Nice, France.

DIWERSY Sascha et VERINE Bertrand (2021) « Observer la capture visuelle en sémantique et en lexicographie », In Lacassain-Lagoin, C., Marsac, F., Ucherek, W. & Chovancova, K. (eds.), *Construction du sens et représentation des référents*. Paris, L'Harmattan, p. 77-94.

GRIES Stefan Th. et HILPERT Martin (2008) « The identification of stages in diachronic data: variability-based neighbor clustering », *Corpora* 3 (1), p. 59-81.

HILPERT Martin et GRIES Stefan Th. (2009) « Assessing frequency changes in multistage diachronic corpora : Applications for historical corpus linguistics and the study of language acquisition », *Literary and Linguistic Computing*, 24 (4), p. 385–401.

LEBART Ludovic et SALEM André (1994) *Statistique textuelle*. Paris, Dunod.

SALEM André (1988) « Approches du temps lexical », *Mots*, 17, p. 105–143. ‹DOI : 10.3406/mots.1988.1401›.

Jeudi 7 novembre, 11h00

L'ancienne sagesse et la naissance des sciences modernes : constitution des corpus à plusieurs niveaux

Eszter Kovács , Cornelis Johannes Schilt, Nicolo Cantoni, Demetrios Paraschos, Jeffrey Wolf (Vrije Universiteit, Brussel)

Le projet ERC VERITRACE (Vrije Universiteit Brussel, porteur du projet : C.J. Schilt, membres d'équipe : N. Cantoni, E. Kovács, D. Paraschos, J. Wolf) étudie l'influence, souvent négligée ou occultée dans l'historiographie des sciences, de la *prisca sapientia* sur la philosophie naturelle durant la première modernité. La période étudiée va de 1540 à 1728, de l'ouvrage *Philosophia perennis* d'Agostino Steuco à la *Chronologie* posthume de Newton, qui croyait encore aux vérités universelles d'origine divine. Le projet cherche à démontrer que l'idée d'une ancienne sagesse communiquée à l'humanité par les dieux et perdue par la suite ne cessa pas de fasciner penseurs et savants même après la découverte philologique d'Isaac Casaubon, qui démontra en 1614 que le *Corpus Hermeticum* avait sans doute été écrit aux II-III^{e} siècles après Jésus-Christ et n'était pas contemporains aux patriarches.

Bien que notre projet n'étudie pas les changements linguistiques dans un corpus diachronique mais l'histoire de certaines idées dans ce type de corpus, les méthodes dont nous nous servons, plus particulièrement les méthodes numériques, sont souvent similaires à celles de la linguistique de corpus. La réflexion sur l'ancienne sagesse pendant la première modernité est caractérisée par un vocabulaire spécifique, par la citation, en latin ou en grec dans un texte en langue vernaculaire, par la paraphrase neutre ou, au contraire, par-

tial. Nous essayons de modéliser l'influence des idées majeures par l'analyse de l'accumulation ou de la régression des éléments lexiques qui caractérisent cette thématique. A part la recherche pour citations et paraphrases, des méthodes comme *Latent Semantic Analysis* and *Sentiment Analysis* nous aident à découvrir des influences cachées. Notre réflexion porte également sur le lien entre l'intertextualité comme phénomène linguistique et littéraire et la notion de l'influence, plus abstraite et difficilement quantifiable.

Pour retracer l'influence de la *prisca sapientia* sur la pensée de la première modernité, voire pour illustrer ses réminiscences tardives à l'époque des Lumières, nous avons établi un corpus des sources (*Close Reading Corpus* ou CRC) et un autre corpus que l'on souhaite exploiter par des méthodes computationnelles (*Distant Reading Corpus* ou DRC). Ce dernier est composé de grands ensembles textuels numérisés susceptibles de contenir des références à l'ancienne sagesse durant la période étudiée.

S'agissant à l'origine de quatre œuvres de longueur différente dont la redécouverte aboutit à un ensemble d'éditions et de traductions, notamment par Pico de la Mirandolla, Ficino, Francesco Patrizi, Thomas Stanley, Jean Le Clerc, notre travail comprend l'analyse des masses textuelles numérisées et rendues accessibles au XXI^{e} siècle. Nous menons des études de cas ponctuelles, tel que l'analyse du champ sémantique relatif au « feu » dans les *Oracles chaldaïques* et dans sa réception, et des analyses computationnelles, utilisant un texte entier du CRC, comparé au DRC. La nature d'un corps textuel soumis à l'analyse, allant d'un seul passage jusqu'à une édition complète, influe sur des méthodes possibles et efficaces.

Le CRC comprend toutes les éditions et traductions connues des *Oracles chaldaïques*, des *Oracles Sibyllins*, des *Hymnes orphiques* et du *Corpus Hermeticum* entre 1540 et 1728. Il s'agit d'un tableau Airtable contenant actuellement 136 entrées. Les éditions recensées dans ce tableau reçoivent un identifiant et sont alignées avec l'identifiant et le permalien USTC, ainsi qu'avec l'URL d'une édition en ligne en libre accès. Un autre type d'indexation, qui prend en compte des noms propres et des lieux d'éditions de ce tableau, sera considéré plus tard. Nous cherchons également à faire des analyses lexicales sur une version OCR de haute qualité de certaines éditions du CRC. Ce travail comprendra des tâches telle que la recognition des entités nommées (NER, outils divers testés) et la recognition des parties du discours (spaCy) pour constituer un vocabulaire de l'ancienne sagesse (*Ancient Wisdom Vocabulary*).

Le DRC comprend les bases textuelles Early English Books Online, Gallica et la collection de Bayerische Staatsbibliothek. Nous avons déjà créé une base spécifique, appelée VEEBO, comprenant les textes d'EEBO pour la période examinée. Concernant Gallica, nous utiliserons l'API du site, et non pas l'inter-

face publique, pour avoir accès à la base textuelle et cibler les requêtes dans les textes et dans les métadonnées.

Dans ma communication, dans un premier temps, je présenterai le travail collectif de l'équipe, qui se concentre sur des corpus à plusieurs niveaux. Je me pencherai dans un deuxième temps sur le lien méthodologique possible entre la recherche en linguistique de corpus et la recherche en histoire des idées. Finalement, j'essayerai d'illustrer par l'étude des syntagmes contenant le mot « feu », provenant des *Oracles Chaldaïques* ou du *Corpus Hermeticum* notre travail comparatif entre le latin, l'anglais et le français.

Références

ALLEN Graham (2000) *Intertextuality*, London, Routledge.

DOBSON James E. (2019) *Critical Digital Humanities : The Search for a Methodology*, Urbana, IL, University of Illinois Press.

HILL Mark J. et HENGCHEN Simon (2019) « Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study », Digital Scholarship in the Humanities, 34, 4, p. 825-843. ‹DOI : 10.1093/llc/fqz024›.

HILPERT Martin et GRIES Stefan Th. (2016) « Quantitative approaches to diachronic corpus linguistics », In Kytö M et Pahta P. (eds.) *The Cambridge Handbook of English Historical Linguistics*, Cambridge, Cambridge University Press, p. 36–53.

LEVITIN Dmitri (2015) *Ancient Wisdom in the Age of the New Science*, Cambridge, Cambridge University Press.

MORETTI Franco (2013) *Distant Reading*, London and Brooklyn, NY, Verso.

SCHMIDT-BIGGEMANN Wilhelm (2004) *Philosophia Perennis: Historical Outlines of Western Spirituality in Ancient, Medieval and Early Modern Thought*, Dordrecht, Springer.

Jeudi 7 novembre, 11h50

L'évolution des mots de l'hétérodoxie dans l'Espagne du XVI^e^ siècle. Éditorialisation outillée et exploration textométrique d'une collection échantillonnée de sources inquisitoriales espagnoles et d'ouvrages de spiritualité

David Kahn (FRAMESPA, INU Champollion, Pau), Sascha Diwersy (PRAXILING, Université de Montpellier 3)

La collection entend interroger la réception en Espagne de la Réforme au XVI^e^ siècle à travers un ensemble représentatif et contrastif de documents inquisitoriaux et de textes de spiritualité. En complément d'un échantillon de procès

de foi impliquant des prévenus soupçonnés ou accusés d'affinités luthériennes, on y trouve un ensemble de sources composé, notamment, de notes de censure, d'édits de la foi et de correspondance inquisitoriale. Et en contrepoint de ces actes judiciaires et administratifs – manuscrits conservés, pour la plupart, à l'*Archivo Histórico Nacional de Madrid* –, la collection réunit 26 livres publiés entre 1528 et 1600, en Espagne et au-delà de ses frontières, et écrits ou traduits par des auteurs ibériques exilés en terre protestante, pour la plupart, et tenus pour des adversaires à combattre aux yeux des autorités inquisitoriales : des ouvrages de catéchèse, des dialogues spirituels, des traités de théologie et des traductions d'autorités patristiques, qui oscillent entre écritures de la dissimulation, visées polémiques et critiques contemptrices.

En 1521, avec la diète de Worms, s'ouvre une période de recomposition confessionnelle où, au fil des radicalisations et des occasions manquées d'une troisième voie consensuelle, sont tracées graduellement les lignes de partage entre orthodoxies et hétérodoxies que consacre le concile de Trente (1547-1563) (Betrán, Hernández et Moreno Martínez 2016, Bœglin 2016). Ainsi des notions et des pratiques communément admises dans la péninsule ibérique, voire promues au début du XVI^{e} siècle, dans le cadre de la politique religieuse impulsée par le cardinal Cisneros (García Oro 1992), deviennent-elles des marqueurs de divergence doctrinale au tournant des années 1520-1530, puis de clivages partisans, propageant des dissensions idéologiques qui, à leur paroxysme, se muent en déchirure de la chrétienté d'Occident. Au cours de ces années où orthodoxie et hétérodoxie sont en travail, des mots-pivots (Guilhaumou 2002), comme *« justificacíon »*, *« líbero arbitri »*, *« fe »*, marquent les évolutions idéologiques, ainsi qu'en attestent les glissements sémantiques et l'émergence des nouvelles acceptions, identifiés par l'analyse des cooccurrences impliquant les mots en question.

C'est la complémentarité typologique de ces documents, produits par des auteurs et des institutions antagonistes, qui permet à l'historien d'opérer des croisements féconds dans le cadre d'approches sociale, culturelle et religieuse. Cependant, à cause de l'hétérogénéité des genres textuels et des déséquilibres dus à une conservation inégale des sources, aux états anciens de la langue et à la variabilité linguistique des formes, une telle collection documentaire implique, d'une part, un traitement outillé, en amont, pour en assurer l'accessibilité : transcription automatique, correction ; indexation et contemporanéisation des formes anciennes afin de parer aux limites des outils de TAL quant à la prise en charge du castillan du XVI^{e} siècle ; application de la chaîne d'annotation, que nous avons construite sur la base de l'analyseur Stanza (Qi *et al.* 2020), aux documents modernisés. Elle suppose, d'autre part, de résoudre des problèmes préalables à l'exploration et à l'analyse quantitative des données textuelles. A cet égard, notre contribution portera sur les défis posés

par l'étude de la charge idéologique que prennent certains mots, à travers l'évolution de leurs co-occurrences, que nous soumettrons à un traitement textométrique au moyen de différents procédés adaptés à la prise en charge de données diachroniques (périodisation automatique par VNC, Gries et Hilpert 2008 ; identification d'accroissements spécifiques à différentes périodes d'une série chronologique, Lebart et Salem 1994, 221-224 ; détection de moments de bascule par Usage Fluctuation Analysis, McEnery et al. 2019).

Références

BETRÁN José Luis, HERNÁNDEZ Bernat et MORENO MARTÍNEZ Doris éd. (2016) *Identidades y fronteras culturales en el mundo ibérico en la Edad Moderna*, Bellaterra, Universitat Autnòma de Barcelona, Servei de Publicacions.

BŒGLIN Michel (2016) *Réforme et dissidence religieuse en Castille au temps de Charles Quint : L'affaire Constantino de la Fuente (1505?-1559)*, Paris, Honoré Champion éditeur.

GARCÍA ORO José (1992) *El cardenal Cisneros : vida y empresas*, Madrid, La Editorial Católica.

GRIES Stefan Th. et HILPERT Martin (2008) « The identification of stages in diachronic data: variability-based neighbor clustering », *Corpora* 3 (1), p. 59-81.

GUILHAUMOU Jacques (2002) « Le corpus en analyse de discours : perspective historique », *Corpus* 1. ‹DOI : 10.4000/corpus.8›.

LEBART Ludovic et SALEM André (1994) *Statistique textuelle*, Paris, Dunod.

QI Peng, ZHANG Yuhao, ZHANG Yuhui, BOLTON Jason et MANNING Christopher D. (2020) « Stanza: A Python Natural Language Processing Toolkit for Many Human Languages », In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

MCENERY Tony, BREZINA Vaclav et BAKER Helen (2019) « Usage Fluctuation Analysis : A new way of analysing shifts in historical discourse », *International Journal of Corpus Linguistics*, 24 (4), p. 413-444.

jeudi 7 novembre, 12h25

Etudier la variation générique en diachronie : le point-virgule comme indice linguistique de différenciation des genres littéraires aux XVIe et XVIIe siècles

Tanguy LEMOINE (Université Grenoble-Alpes)

Cette communication aura pour but de présenter les enjeux de la constitution et de l'exploitation d'un corpus permettant de mêler les études littéraires et linguistiques de l'évolution des genres narratifs de la prose française des XVIe

et XVII[e] siècles, dans une perspective de stylistique des genres (Combe 2002) et de linguistique textuelle diachroniques (Combettes 2012, 2023). Il s'agira de mettre en évidence l'émergence du point-virgule et d'analyser ses différentes fonctions au sein de trois grands genres de la prose narrative préclassique : le roman, le récit bref et le récit de voyage.

L'enjeu linguistique de l'apparition du point-virgule, au niveau de la structuration textuelle, a déjà été en partie exploré (Dürrenmatt 2011, Lorenceau 1984), mais sans toujours bénéficier des méthodes quantitatives applicables à un corpus qui prenne en compte à la fois de l'instabilité de la ponctuation inhérente à l'histoire de l'imprimé durant cette période (Catach 1977, McKitterick 2018, Dourdy et Spacagno 2021) et des différenciations d'usage des signes de ponctuation au sein de genres textuels variés. Les signes de ponctuation intermédiaires constituent pourtant des lieux cruciaux, mais encore peu exploités, pour l'exploration des dynamiques diachroniques et génériques de structuration textuelle (Goux 2021, Ström Herold et Levin 2023).

L'hypothèse testée est que l'analyse des fonctions linguistiques du point-virgule et de son importance dans l'organisation textuelle permet d'ouvrir la voie à une meilleure caractérisation littéraire des genres (perspective déjà ouverte récemment au niveau de l'histoire de la phrase française (Siouffi *et al.* 2020, Mounier 2022), dans une approche non quantitative).

Cette démarche a nécessité la constitution d'un corpus spécifique, qui soit rigoureusement fidèle au niveau de la transcription des signes de ponctuation de l'imprimé source et enrichi par une annotation de la structure textuelle des échantillons, tout en exploitant et valorisant au mieux les corpus déjà existants pour la période d'étude. Ces derniers ont été précieux dans la construction du corpus, mais aucun n'offrait cependant une exhaustivité générique suffisante (Frantext[1], PhraséoRoChe[2]) ou ne couvrait tout l'empan diachronique du français préclassique (Bibliothèques Virtuelles Humanistes[3], Corpus Electroniques de la Première Modernité[4]). Il a alors été nécessaire d'utiliser des textes provenant de corpus variés, voire issus de simples éditions numérisées ou des transcriptions propres. Seront donc abordées les problématiques de constitution d'un corpus à partir de sources hétérogènes et les solutions mises en place pour proposer à moyen terme un corpus exploitable par la communauté scientifique (Denoyelle *et al.* 2024).

Enfin, cette communication sera l'occasion d'aborder des outils d'exploitation du corpus, notamment dans une double perspective linguistique et littéraire, qui nécessite la possibilité d'un constant retour au texte (possibilité notamment offerte par le logiciel TXM (Heiden, Magué et Pincemin 2010)). Il s'agira également de présenter des méthodes d'analyse qui permettent de rendre compte des évolutions à la fois générique et diachronique, de révéler les

variations internes aux genres et aux périodes, mais aussi les dynamiques de rapprochement et de différenciation.

1. https://www.frantext.fr
2. https://lidilem.univ-grenoble-alpes.fr/node/16/axes-recherche/axe-1-description-et-modelisation-linguistiques-corpus-tal/projets-laxe-1/phraseo-13-18
3. https://www.bvh.univ-tours.fr
4. http://www.cepm.paris-sorbonne.fr

Références

CATACH Nina (1977) « La ponctuation dans les imprimés, des débuts de l'imprimerie à G. Tory et E. Dolet », In *La Ponctuation. Recherches historiques et actuelles*, Paris / Besançon, CNRS-GTM-HESO, p. 29-57.

COMBE Dominique (2002) « La stylistique des genres », *Langue française*, n° 135, p. 33-49.

COMBETTES Bernard (2012) « Linguistique textuelle et diachronie », In *Actes du 3e Congrès Mondial de Linguistique Française*, Lyon, EDP Sciences, p. 3-10.

COMBETTES Bernard (2023) « Suggestions for a diachronic text linguistics », In *French theories on text and discourse*, Beihefte zur Zeitschrift für romanische Philologie, Berlin, Boston, De Gruyter, p. 169-184. ‹DOI : 10.1515/9783110794434-009›.

DENOYELLE Corinne, KRAIF Olivier, MOUNIER Pascale, *et al.* (2024) « Le corpus PhraséoRoChe : les défis de l'établissement des textes et de l'hétérogénéité des états de la langue », *Corpus*, n° 25. ‹DOI : 10.4000/corpus.8501›.

DOURDY Laura-Maï et SPACAGNO Michela (2021) « Variance typographique et évolution linguistique : analyse de la ponctuation dans cinq traditions textuelles imprimées au XVIe siècle », *Çédille, revista de estudios franceses*, n° 19, p. 227-255. ‹DOI : 10.25145/j.cedille.2021.19.10›.

DÜRRENMATT Jacques (2011) « Grandeur et décadence du point-virgule », *Langue française*, n° 172, p. 37-52.

GOUX Mathieu (2021) « Ponctuation et connecteurs en français classique. Du reposoir (périodique) à la structure (phrastique) », *Çédille, revista de estudios franceses*, n° 19, p. 127-156. ‹DOI : 10.25145/j.cedille.2021.19.06›.

HEIDEN Serge, MAGUÉ Jean-Philippe et PINCEMIN Bénédicte (2010) « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », In *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*, Rome, LED Edizioni Universitarie, p. 1021-1032.

LORENCEAU Annette (1984) « Histoire du point-virgule et des deux points dans la ponctuation française », In *Trames. Actualité de l'histoire de la langue*

française, Limoges, Travaux et Mémoires de l'Université de Limoges, p. 99-107.

McKitterick David (2018) *Textes imprimés et textes manuscrits. La quête de l'ordre, 1450-1830*, Lyon, ENS Éditions / Institut d'histoire du livre.

Mounier Pascale (2022) « La phrase dans le roman de chevalerie entre 1530 et 1550 », *Le Français préclassique*, n° 24, p. 97-123.

Siouffi Gilles, Marchello-Nizia Christiane, Combettes Bernard, *et al.* (2020) *Une histoire de la phrase française des Serments de Strasbourg aux écritures numériques*, Arles, Actes Sud / Imprimerie nationale Éditions.

Ström Herold Jenny et Levin Magnus (2023) « Tracing ongoing change in Swedish punctuation – a corpus-based study », Regensburg, "Comparative Punctuation Worldwide" conference.

jeudi 7 novembre, 14h15

Constitution d'un corpus longitudinal par la transcription automatique (HTR/OCR) : problèmes liés à l'hétérogénéité des sources

Denisa Bumba (Université Sorbonne nouvelle), Gabriella Parussa (Sorbonne Université, STIH)

En nous plaçant dans la perspective du traitement des sources utilisées dans la constitution d'un corpus textuel diachronique et numérique, nous souhaitons rendre compte des difficultés spécifiques que pose la transcription automatique (HTR/OCR) par le biais des outils du type E-scriptorium, Transkribus ou OCR4all. En effet, dans le cadre de l'ANR EcoLe, lors de la constitution d'un corpus textuel longitudinal (13^{e} – 18^{e} s.) nous avons soit dû transcrire des sources manuscrites et imprimés soit intégrer des textes déjà transcrits et encodés pour d'autres projets. L'hétérogénéité de ces sources a constitué d'emblée un défi majeur aussi bien pour l'HTR/OCR que pour le traitement des données textuelles dans le but de constituer un corpus étiqueté et exploitable par des outils de textométrie.

Par cette contribution, nous voudrions faire état des problèmes rencontrés et des solutions trouvées pour l'application de l'HTR à nos sources (évaluation et choix du modèle, fine-tuning de modèle, principes de transcription, traitement et uniformisation des résultats de l'HTR/OCR). La mise en page étant une donnée importante pour notre projet, nous avons dû chercher des moyens pour ajouter ces indications sous forme de balisage à la transcription du texte. La détection des colonnes, des marginalia, des illustrations et des décorations a donc constitué un des premiers défis dans la production de transcriptions. Ensuite, la variété des écritures a exigé l'évaluation de plusieurs modèles HTR/OCR existants, avec souvent une étape de réentraînement de modèle.

Tiraillées entre les exigences philologiques (fournir une transcription fiable et fidèle, respecter les graphies, la segmentation et la mise en page originales) et les contraintes imposées par le traitement automatique ultérieur du corpus (comme la lemmatisation, l'identification des entités nommées etc.), les linguistes doivent faire face à de véritables dilemmes : faut-il viser l'efficacité, la rapidité en admettant un taux d'erreurs important ou bien réduire la taille du corpus afin d'obtenir la plus haute fiabilité possible des transcriptions et un résultat recevable de l'étiquetage pour permettre des fouilles textuelles de meilleure qualité ?

De plus, l'empan chronologique assez vaste du corpus diachronique en question, a créé des disparités entre les modalités de transcription, dues aussi bien à des pratiques divergentes entre les médiévistes et les dix-septièmistes, par exemple, qu'à la variabilité des supports, des écritures et du code écrit.

Références

REUL Christian, CHRIST Dennis, HARTELT Alexander, BALBACH Nico, WEHNER Maximilian, SPRINGMANN Uwe, Wick Christoph, Grundig Christine, Büttner Andreas, et Puppe Frank (2019) « OCR4all—An open-source tool providing a (semi-)automatic OCR workflow for historical printings ». *Applied Sciences* 9, n° 22 : 4853.

VIDAL-GORÈNE Chahan, DUPIN Boris, DECOURS-PEREZ Aliénor, et RICCIOLI Thomas (2021) « A modular and automated annotation platform for handwritings : evaluation on under-resourced languages ». *International Conference on Document Analysis and Recognition - ICDAR 2021*, dir. LLADÓS Josep, LOPRESTI Daniel, et UCHIDA Seiichi. Lecture Notes in Computer Science, vol. 12823. Cham : Springer, 507-522. https://doi.org/10.1007/978-3-030-86334-0_33.

CAMPS Jean-Baptiste, VIDAL-GORÈNE Chahan, STUTZMANN Dominique, VERNET Marguerite, et PINCHE Ariane (2022) « Data Diversity in handwritten text recognition, Challenge or opportunity? », prés. *Digital Humanities 2022*, Tokyo, 27 juillet 2022.

RUSCIO Maxime, ROILAND Muriel, MALOBERTI Sarah, NOËMIE Lucas, PERRIER Antoine, et VIDAL-GORÈNE Chahan (2022) « Les collections de manuscrits maghrébins en France (2/2) » . HAL, https://medihal.archives-ouvertes.fr/hal-03660889.

CAMPS Jean-Baptiste, VIDAL-GORÈNE Chahan et VERNET Marguerite (2021) « Handling Heavily Abbreviated Manuscripts: HTR engines vs text normalisation approaches ». In *International Conference on Document Analysis and Recognition - ICDAR 2021*, dir. Elisa H. Barney Smith, Umapada Pal. Lecture Notes in Computer Science, vol. 12917. Cham : Springer, 507-522. https://doi.org/10.1007/978-3-030-86159-9_21.

TORRES AGUILAR Sergio, JOLIVET Vincent (2023) « Handwritten Text Recognition for Documentary Medieval Manuscripts » *Journal of Data Mining and Digital Humanities (Historical Documents and automatic text recognition).* https://hal.science/hal-03892163.

jeudi 7 novembre, 14h50

D'une encyclopédie à l'autre: le corpus «parallèle» du projet GEODE

Alice Brenon (Laboratoire d'InfoRmatique en Image et Systèmes d'information / LIRIS) - Institut National des Sciences Appliquées de Lyon & Interactions, Corpus, Apprentissages, Représentations / ICAR -Université Lumière - Lyon 2)

Publiée au tournant du XIX^e^ et du XX^e^ siècle (de 1885 à 1902), La Grande Encyclopédie, Inventaire raisonné des Sciences, des Lettres et des Arts par une Société de savants et de gens *de lettres* (*LGE*) est d'après Jacquet-Pfau (2015, 85) la dernière grande entreprise encyclopédique française à marcher dans les traces de l'*Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres*, dite « de Diderot et d'Alembert » (*EDdA*) qu'elle évoque jusque dans son titre complet.

Ces deux encyclopédies figurent dans le corpus d'étude de GEODE[1], un projet pluridisciplinaire qui observe les changements survenus dans les discours géographiques au sein des encyclopédies françaises depuis le XVIII^e^ siècle. Des travaux diachroniques menés dans le cadre du projet ont bien sûr exploré les changements de thématiques abordées, mais une des études a consisté au contraire à étudier les variations des discours portant sur un ensemble fixe d'objets entre les deux œuvres. C'est l'origine d'un corpus dit « parallèle » qui met en regard un ensemble de 3 706 articles de l'*EDdA* avec leurs équivalents dans *LGE* (l'ensemble du corpus contient donc 7 412 articles au total).

Cette communication qui s'inscrit dans le premier axe « Traitement des corpus numériques diachroniques », met en jeu des données issues de sources différentes : l'ARTFL[2] pour la moitié EDdA du corpus parallèle (Morrissey et Roe 2022), une version numérisée et OCRisée par la BnF issue du projet DISCO-LGE[3] (Vigier et Brenon 2021) pour la moitié *LGE*. Elle sera l'occasion de présenter leurs différences et de montrer les défis posés par leur hétérogénéité pour la constitution du corpus parallèle. D'un point de vue méthodologique, elle abordera également quelques aspects techniques de la définition de ce corpus : la description de l'algorithme suivi pour parvenir à cet ensemble précis d'articles encyclopédiques, les méthodes utilisées pour évaluer sa qualité ainsi que les choix opérés en termes de métadonnées pour refléter la structure particulière de ce corpus.
La focalisation du projet GEODE sur la Géographie amène naturellement à partitionner les articles par discipline scientifique et à s'intéresser aux mé-

thodes pour les classer automatiquement (Brenon et al. 2022). Cette approche quand elle est confrontée à un objet comme le corpus parallèle soulève plusieurs problèmes, notamment épistémologiques, pour la détermination d'un ensemble de domaines de connaissances comparables d'une moitié à l'autre du corpus mais pertinents à chacune des époques. Ces difficultés seront discutées en pratique dans une étude sur des entrées qui ont changé de domaine entre l'*EDdA* et *LGE*.

1. https://geode-project.github.io
2. https://artfl-project.uchicago.edu
3. https://www.collexpersee.eu/projet/disco-lge

jeudi 7 novembre, 15h25

La Dicothèque : un outil pour l'exploration pluridimensionnelle des dictionnaires

Lucas Lévêque, Florian Cuny, Noé Gasparini (Université de Lyon3, Institut international pour la Francophonie)

Avec le développement des outils numériques sont apparues de nouvelles possibilités de compilation des connaissances, qui répondent à des besoins documentaires, notamment pour la réalisation de dictionnaires, grâce à l'accès rapide à des ouvrages lexicographiques. C'est ainsi que l'outil libre Dicothèque (https://dicotheque.org) a été développé au sein de la communauté Wikimédia. Il est à l'interface de trois de ses projets principaux : Wiktionnaire, Wikisource et Wikidata.

Il est né de l'observation des besoins liés à la pratique de la lexicographie collaborative au sein du Wiktionnaire francophone. Le contributorat s'appuie sur des travaux existants, et notamment des dictionnaires entrés dans le domaine public, dont le contenu est plus facilement accessible. Ce besoin va au-delà, puisqu'il est utile pour des recherches aussi bien que pour une consultation par le grand public.

Dans le même temps, le projet collaboratif Wikisource s'est développé, proposant des ouvrages relus à partir de fac-similés, publiés au fil des relectures, pages à pages. Il intègre un nombre croissant d'ouvrages lexicographiques hétérogènes, allant de 1606 à 1941, devenant une source de connaissance majeure pour la lexicographie. Une méthodologie a alors été établie pour la construction d'un outil dont le contenu puisse être enrichissable et améliorable en continu.

Cette approche technique est nourrie de l'analyse du fonctionnement des grands outils polylexicaux comme Corpus DiCo, Dictionnaires d'autrefois, Métadictionnaires, Nénufar. Elle est consolidée par les métadonnées des ouvrages décrits dans la base de connaissances Wikidata. Leur description fine

permet un filtrage précis sur les ouvrages consultés, selon la nature des dictionnaires, leurs époques, leurs thématiques ou encore leurs auteurs.

De cette synergie entre projets numériques collaboratifs naissent de nouvelles perspectives : analyses quantitatives, publications multilingues, connexion ou intégration de ressources extérieures, usages pédagogiques. De sa conception à ses usages, les grandes lignes seront tracées pour inscrire la Dicothèque au sein des grands corpus numériques diachroniques.

Références

BOHBOT Hervé, FRONTINI Francesca, KHAN Fahad, KHEMAKHEM Mohamed et ROMARY Laurent (2019) « Nénufar : Modelling a Diachronic Collection of Dictionary Editions as a Computational Lexical Resource », In *ELEX 2019*. ‹URl : https://inria.hal.science/hal-02272978›

CHAMBAT Anais, ROUSSEAU Nathalie et VINCENT Jean-Francois (2022) « Faire dialoguer les anciens dictionnaires de médecine à l'ère du numérique », In *12e* Conférence *Internationale de Lexicographie Historique et de Lexicologie (ICHLL)*, Lorient.

MARTINEZ Camille (2013) « La comparaison de dictionnaires comme méthode d'investigation lexicographique », *Lexique*, vol. 21, p. 193-220.

SAJOUS Franck et MARTINEZ Camille (2022) « Metalexicographical Investigations with the DiCo Database », *International Journal of Lexicography*, vol. 35, n° 1, p. 75-106. ‹DOI : 10.1093/ijl/ecab017›.

TITTEL Sabine (2010) « Dynamic access to a static dictionary: a lexicographical "cathedral" lives to see the twenty-first-century - the "Dictionnaire étymologique de l'ancien français" », In *E Lexicography in the 21st Century New Challenges New Applications. Proceedings of Elex 2009*, Louvain La Neuve.

YOAKIM William (2019) « Wikipédia, Wikimedia Commons et Wikisource, un eldorado de visibilité », *Archives*, vol. 48, n° 2, p. 41-81. ‹DOI : 10.7202/1067524ar›.

Jeudi 7 novembre, 16h30

L'IA peut-elle vraiment nous aider à explorer les grands corpus littéraires ?

Thierry POIBEAU (Langues, Textes, Traitements informatiques, Cognition - Lattice, CNRS)

L'intelligence artificielle est de plus en plus employée pour l'analyse de grands corpus littéraires qui, de fait, dépassent de par leur taille ce qu'il est possible d'appréhender directement (tout simplement parce qu'il est humainement

impossible de lire tous ces textes, par exemple les 15.000 romans du 19e siècle disponibles à travers le site Gallica de la BnF). Mais qu'apprend-on vraiment ainsi ? La question est débattue, entre ceux qui pensent qu'on fait face à une révolution, et ceux qui défendent l'opinion que l'IA ne fait que confirmer des éléments déjà connus, voire n'a aucun intérêt, car seule une connaissance intime des textes permet de les interpréter. Au cours de cet exposé, je partirai d'expériences récentes menées au sein du laboratoire Lattice. Je discuterai les résultats obtenus : en quoi ils peuvent être intéressants (et peut-être apporter des éléments d'analyse nouveaux), et en quoi ils demeurent limités et perfectibles. Je discuterai également du lien entre ces techniques et le besoin de lecture proche et de connaissance fine des textes, qui ne doivent pas être oubliés.

Vendredi 8 novembre, 9h30

Le projet Open French Corpus : constituer un corpus à partir de l'existant

Céline POUDAT (Bases, corpus, langage - BCL, Université Côte d'Azur)

Je présenterai les objectifs et les avancées du projet *Open French Corpus* du consortium CORLI, qui vise à regrouper les corpus existants sous une forme standardisée, tant au niveau des formats que des métadonnées. L'initiative propose de centraliser ces corpus dans un espace commun accessible, accompagné d'outils spécifiques pour leur exploitation. L'objectif est d'améliorer la cohérence, l'accès et l'utilisation de ces ressources, tout en garantissant la qualité des données à travers un processus de validation par la communauté scientifique. Ce projet répond ainsi aux besoins croissants d'une utilisation harmonisée et fiable des corpus dans divers domaines de recherche.

Vendredi 8 novembre, 10h30

La constitution d'un corpus pour l'analyse des disparitions lexicales (XIIIe-XVe siècles) : approches computationnelles et qualitatives

Lucence ING (Centre Jean Mabillon, École nationale des chartes)

Dans notre communication, nous souhaiterions présenter les travaux sur les disparitions lexicales en diachronie que nous avons menés dans le cadre de notre thèse de doctorat, et leurs prolongements. Notre sujet d'étude porte sur les disparitions de mots survenues entre le début du XIIIe siècle et la fin du XVe siècle, soit à plus de deux siècles d'intervalle, entre les états de langue « ancien » et « moyen » français. Elle repose en premier lieu sur l'étude de l'évolution du lexique au sein de la tradition d'une œuvre particulière, le *Lancelot en prose*[1].

Nous nous proposons de présenter comment nous avons constitué les données de notre corpus en fonction de cet objectif d'étude, de la récupération

du texte des témoins à l'extraction automatique des mots disparus, en passant par des étapes automatiques d'annotation linguistique et d'alignement des témoins (faisant particulièrement appel à un modèle d'annotation linguistique fonctionnant sur de l'apprentissage profond, le modèle *pie* pour l'ancien français, pour la première étape, et à un module de collation automatique de témoins, Collatex, pour la seconde).

Du fait de la variabilité des traditions textuelles littéraires médiévales, et de la multiplicité des causes de celle-là, la détection automatique doit être validée par une vérification systématique des mots identifiés comme disparus, nécessitant des connaissances philologiques et linguistiques.

Par ailleurs, la compréhension de l'évolution du lexique implique l'exploration de sources lexicographiques et de sources textuelles extérieures au corpus, et leur interprétation, ce qui est également un travail qui n'est pas réalisé automatiquement.

Nous présenterons donc les divers aspects de la constitution de notre corpus et de ses analyses, au sein desquelles quantitatif et qualitatif sont étroitement mêlés et se révèlent aussi indispensables l'un que l'autre à l'appréhension de notre objet d'étude.

Les prolongements de notre recherche consistent notamment en la récupération d'un plus grand nombre de données textuelles, celles du *Tristan en prose*[2], permettant la comparaison de l'évolution du lexique au sein du *Lancelot* avec celle survenue au sein de la tradition de cette autre œuvre, nouvelles recherches que nous présenterons. Si l'ajout de données permet de confronter une partie des résultats obtenus, nous interrogerons aussi la faisabilité d'une recherche basée sur la constitution de tels corpus, cette dernière étant extrêmement chronophage.

1. L'étude se base principalement sur la comparaison d'un témoin du premier tiers du XIII^e^ siècle, le manuscrit BnF français 768, d'après l'édition de E. Kennedy, et d'un exemplaire de l'*editio princeps* du *Lancelot,* imprimé par Jehan le Bourgois en 1488.
2. Le *Tristan* a été imprimé pour la première fois, par Jehan le Bourgois, en 1489, donc par le même imprimeur que le *Lancelot*. La version du *Tristan* utilisée par l'imprimeur est connue. Il s'agit de celle, abrégée, du manuscrit BnF, français 103, une sous-version de la version IV, qui a comme particularité de proposer une fin proche de celles des versions en vers. Cette version a été composée entre le milieu du XIV^e^ siècle et le milieu du XV^e^ siècle et le manuscrit est daté du troisième quart du XV^e^ siècle ; l'étude sur le *Tristan* semble donc *a priori* porter sur un état de langue plus récent que celui dans lequel a été composé le *Lancelot*, ce qui n'exclut pas les influences linguistiques antérieures, mais complexifie encore la question. Le recours a un témoin antérieur du texte, le manuscrit 2542 de la Bibliothèque nationale de Vienne, daté du début du XIV^e^ siècle et en partie édité, sur des passages comparables, permettra d'explorer plus en avant la question.

Références

BADIOU-MONFERRAN Claire (2008) « Les disparitions de formes sont-elles des épiphénomènes ? » In *Congrès mondial de linguistique française 2008*, p. 147-58, 2008. ‹DOI : 10.1051/cmlf08296›.

BADIOU-MONFERRAN Claire et VERJANS Thomas, éd. (2015) *Disparitions. Contributions à l'étude du changement linguistique,* Paris, Honoré Champion Éditeur.

CAPIN Daniéla (2004) « Le conservatisme de la langue, gage du caractère littéraire du texte et témoin d'une nouvelle conception de l'acte d'écriture : le cas d'Isaÿe le triste », *Medium Ævum* 73, 1, p. 66-92. ‹DOI : 10.2307/43630699›.

DE CARNÉ Damien et FERLAMPIN-ACHER Christine, éd. (2021) *La Tradition manuscrite du Tristan en prose: bilan et perspectives*, Paris, Classiques Garnier Numérique.

KENNEDY Elspeth, éd. (1980) *Lancelot Do Lac : The Non-Cyclic Old French Prose Romance*, Oxford/New York, Clarendon Press.

WINN Mary Beth, éd. (2020) *Tristan, chevalier de la Table Ronde. Tome I : Roman imprimé en 1489 par Jehan Le Bourgois pour Anthoine Vérard*, Paris, Classiques Garnier Numérique. ‹DOI : 10.15122/ISBN.978-2-406-09518-7›.

Outils mentionnés

CollateX, outil de collation automatique, module python. http://interedition.github.io/collatex/pythonport.html.
L'ensemble des scripts pour l'alignement et la collation automatiques est disponible sur https://github.com/LucenceIng/alignementEtCollation.

Pie : MANJAVACAS Enrique, KÁDÁR Ákos et KESTEMONT Mike (2019) « Improving lemmatization of non-standard languages with joint learning ». In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies. Volume 1 : long and short papers*, 1493–1503. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. ‹DOI : 10.18653/v1/N19-1153›.

Vendredi 8 novembre, 11h30

Comparaison de deux approches pour l'analyse syntaxique du français et du latin en diachronie

Mathieu Dehouck, Sophie Prévost (Langues, Textes, Traitements informatiques, Cognition / Lattice, CNRS), Mathilde Regnault (Université de Stuttgart), Loïc Grobol (Modèles, dynamiques, Corpus / MoDyCo, Université Paris Nanterre)

Nous nous intéressons à l'évolution du français, en particulier dans la période médiévale (ancien français 9e-13e s. et moyen français 14e-15e s.), et à son interaction avec le latin médiéval, largement utilisé en Europe comme langue de l'écrit et de diffusion des idées. Rassembler ces corpus pour entraîner des analyseurs syntaxiques et les évaluer pose un premier défi d'interopérabilité des données.

Pour le français médiéval, nous disposons du corpus arboré *Profiterole* (Prévost et al. 2023), divisé en deux parties. La première, UD_Old_French-PROFITEROLE, est constituée de textes d'ancien français et a été annotée dans un premier for- mat dépendanciel créé spécifiquement pour cette ressource, puis converti semi-automatiquement au format *Universal Dependencies* (UD, Zeman et al. 2024).

La deuxième partie, UD_Middle_French-PROFITEROLE, concerne le moyen français et a été directement annotée au format UD. La gestion de cette ressource est contrainte par le manque de spécialistes employés à cette tâche, alors qu'il est recommandé de procéder en équipe (Fort 2016). Il reste des corrections à apporter aux transformations automatiques, dont les erreurs sont parfois difficiles à identifier, et l'adaptation aux évolutions des règles UD demande une veille régulière, en particulier pour des états de langue anciens (cf. initiative UD4HL). Le défi soulevé par la deuxième période de ce corpus est de suivre les mêmes règles d'annotation, mais pour un état de langue plus récent.

En parallèle, UD dispose de textes latins allant de la période classique (Celano 2019) à la période médiévale : *Late Latin Charter Treebank* (LLCT, 774-897, Korkiakangas 2021), *Index Thomisticus* (ITTB, 1225–1274) et UDante (14e s.) (Gamba et Zeman 2023).

Par le biais d'expériences d'analyse syntaxique, nous souhaitons évaluer l'interopérabilité de ces données pour mettre au point de meilleurs outils d'annotation. Ceux-ci permettront ensuite d'évaluer l'influence du latin médiéval dans les productions francophones et inversement. Nous souhaitons aussi apporter des réponses à la question suivante : s'il est généralement vrai que plus de données conduisent à de meilleurs résultats dans quelle mesure cela s'applique-t-il à une diachronie longue ?

Nos premières expériences sont organisées en deux axes. Dans le premier, nous cherchons avant tout à répondre à notre question en entraînant des modèles d'analyse syntaxique sur de vastes empans chronologiques. Les premiers résultats indiquent que l'apport d'états de langue divers est positif, y compris pour le français contemporain, pourtant mieux doté et plus normé. Dans le deuxième axe, nous entraînons des modèles d'analyse syntaxique sur

des empans plus courts et les évaluons sur des textes d'autres époques, de la même langue et d'autres langues. Les premiers résultats montrent que les modèles d'analyse sont plus performants quand appliqués à des textes d'époques et de genres proches de leurs données d'entraînement dans le cadre monolingue. Il s'agit maintenant d'étendre cette étude au cas multilingue.

Références

CELANO, Giuseppe G.A. (2019) « The Dependency Treebanks for Ancient Greek and Latin ». In : *Ancient Greek and Latin in the Digital Revolution*. Sous la dir. de Monica Berti. Berlin, Boston : De Gruyter Saur, p. 279-298. ‹DOI : 10.1515/9783110599572-016›.

FORT, Karën (2016) Collaborative Annotation for Reliable Natural Language Processing : Technical and Sociological Aspects. Sous la dir. de Patrick Paroubek. Wiley-ISTE. ‹URL : https://hal.science/hal-01324322›.

GAMBA, Federica , ZEMAN Daniel (2023) « Universalising Latin Universal Dependencies : a harmonisation of Latin treebanks in UD ». In : *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*. ‹URL : https://aclanthology.org/2023.udw-1.2›.

KORKIAKANGAS, Timo (2021) « Late Latin Charter Treebank : contents and annotation », *Corpora* 16.2, p. 191-203. ‹DOI : 10.3366/cor.2021.0217›.

PRÉVOST, Sophie *et al.* (2024) « Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval », *Corpus* 25. ‹DOI : 10.4000/corpus.8538›.

ZEMAN, Daniel et al. (2024) *Universal Dependencies 2.14*. LINDAT/CLARIAH- CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. ‹URL : http:// hdl.handle.net/11234/1-5502›.

Vendredi 8 novembre, 12h05

Quelques pistes pour surmonter les contraintes pour l'annotation syntaxique de corpus en diachronie longue

Natasha Romanova, Rayan Ziane (CRISCO, Université de Caen Normandie)

S'inscrivant dans l'axe 1 du colloque, la communication proposée sera consacrée aux défis de l'annotation syntaxique de corpus en diachronie longue.

L'émergence de gros modèles de langue (LLMs) et la mise à disposition récentes de plusieurs outils neuronaux d'analyse syntaxique en dépendances (*parsers*), a considérablement facilité la tâche de l'annotation syntaxique (en catégories et en fonctions) et l'outillage pour la recherche en syntaxe sur corpus (Straka et al. 2016, Guiller 2020, Grobol et Crabbé 2021). La qualité de

l'annotation automatique dépendant en grande partie de la distance entre le corpus d'entraînement et le corpus cible, la précision de l'annotation d'un nouveau corpus peut être améliorée soit par un programme de correction en post-traitement soit via *bootstrapping* qui un processus d'adaptation progressive du modèle au texte cible (Peng et al. 2022).

Lors du travail sur un corpus en français en diachronie longue (12^{e}-17^{e} siècles), calibré par genre (chroniques), nous avons identifié deux verrous importants. Le premier groupe de défis (Verrou 1) est lié à la disponibilité de corpus de référence pour l'entraînement de modèles ainsi que la distance entre les corpus d'entraînement et les corpus cible, notamment en ce qui concerne la variation diachronique et de genre. Sur le site du projet *Universal Dependencies* (UD; de Marneffe et al. 2021) les corpus ne sont disponibles que pour la période médiévale (corpus Profiterole, ancien corpus SRCMF; Prévost et Stein 2013; Prévost et al. 2024) et la fin du 20^{e} - début du 21^{e} siècles (corpus Sequoia; Candito et Seddah 2012). Aucune ressource validée n'existe jusqu'ici pour les périodes entre le 16^{e} et le 19^{e} siècles.

Le deuxième type de contrainte (Verrou 2) est lié au fait que au sein du projet UD l'ancien français et le français moderne sont considérés comme des langues différentes avec certaines incohérences entre les principes d'annotation, notamment le statut des verbes modaux (considérés comme auxiliaires dans SRCMF/Profiterole et comme tête de la proposition dans les corpus du français moderne), ce qui aurait un impact sur l'évaluation ou le réentraînement de systèmes de parsing qui porteraient sur les états de la langue qui s'étendent entre le Moyen Âge et la période moderne tout en entravant l'analyse comparée des textes différentes époques.

Afin de pallier aux incohérences d'annotation, nous avons commencé par aligner l'annotation du corpus Profiterole aux principes de l'annotation du français moderne pour le rendre compatible avec le corpus Sequoia. Nous avons ensuite procédé à une série de tests sur des échantillons des textes du 13^{e} et 17^{e} siècle de notre corpus dont nous disposons des versions « Gold » vérifiées par l'équipe pour trouver les systèmes de parsing le mieux adaptés à l'annotation d'un corpus en diachronie longue.

Les résultats de ces expériences seront présentés dans cette intervention afin de stimuler la discussion dans la communauté qui permettrait de trouver des solutions d'interopérabilité de systèmes d'annotation pour les différents états de la langue (Verrou 2) ce qui, par conséquent, faciliterait l'émergence de corpus de référence pour les périodes encore peu dotées de l'histoire du français (Verrou 1).

Références

CANDITO Marie et SEDDAH Djamé (2012) « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical », *Actes de TALN 2012. Grenoble.* 15 p.

DE MARNEFFE Marie-Catherine, MANNIING Christopher D., NIVRE Joakim et ZEMAN Daniel (2021) « Universal Dependencies », Computational Linguistics, vol. 47, n° 2, pp. 255-308.

GROBOL Loïc et CRABBÉ Benoît (2021) « Analyse en dépendances du français avec des plongements contextualisés (French dependency parsing with contextualized embeddings) », *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*, vol. 1 : conférence principale, pp. 106-114.

GUILLER Kirian (2020) *Analyse syntaxique automatique du pidgin-créole du Nigeria à l'aide d'un transformer (BERT): Méthodes et Résultats*. (Mémoire de Master, Sorbonne Nouvelle)

PENG Ziqian, GERDES Kim et GUILLER Kirian (2022) « Pull your treebank up by its own bootstraps », *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, Marseille, pp.139-153.

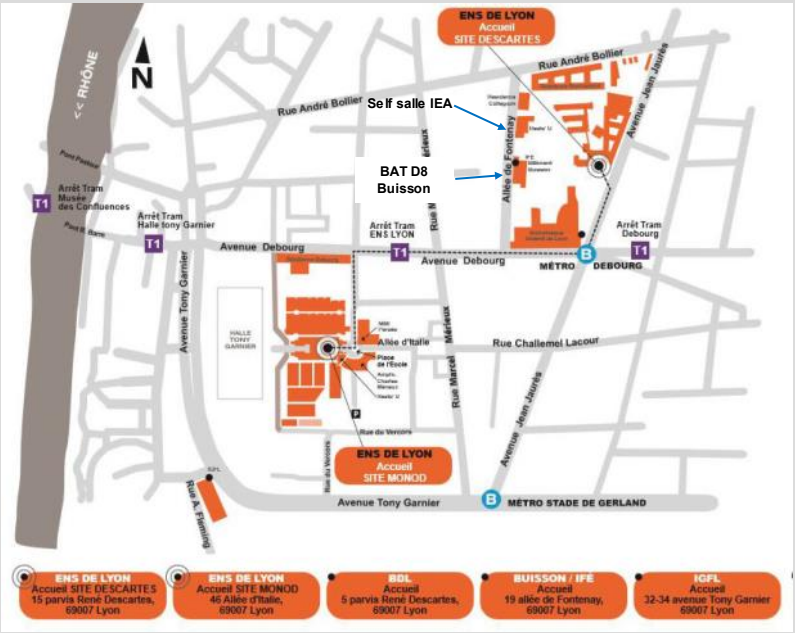
PRÉVOST Sophie et STEIN Achim (2013) « Syntactic annotation of medieval texts: the Syntactic Reference Corpus of Medieval French (SRCMF) », In P. Bennett, M. Durrell, S. Scheible et R. Whitt (éd.), *New Methods in Historical Corpus Linguistics*, Tübingen, Narr Verlag, pp. 275-282.

PRÉVOST Sophie, GROBOL Loïc, DEHOUCK Mathieu, LAVRENTIEV Alexei et HEIDEN Serge (2024) « Profiterole : Un corpus morpho-syntaxique et syntaxique de français médiéval » *Corpus*, vol. 25. ‹DOI : 10.4000/corpus.8538›.

STRAKA Milan, HAJIČ Jan et STRAKOVÁ Jana (2016) « UDPipe : Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing », *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, Slovenia. p. 4290–4297.

Comité d'organisation

Céline Guillot-Barbance, (ENS de Lyon, IHRIM, France)
Alexei Lavrentiev, (ENS de Lyon, IHRIM, France)
Tanguy Lemoine, (Université de Grenoble Alpes, Litt&Arts)
Raphaël Luis, (ENS de Lyon, CERCC, France)
Matthieu Quignard, (ICAR, CNRS, France)



CONTACT

concordial@sciencesconf.org