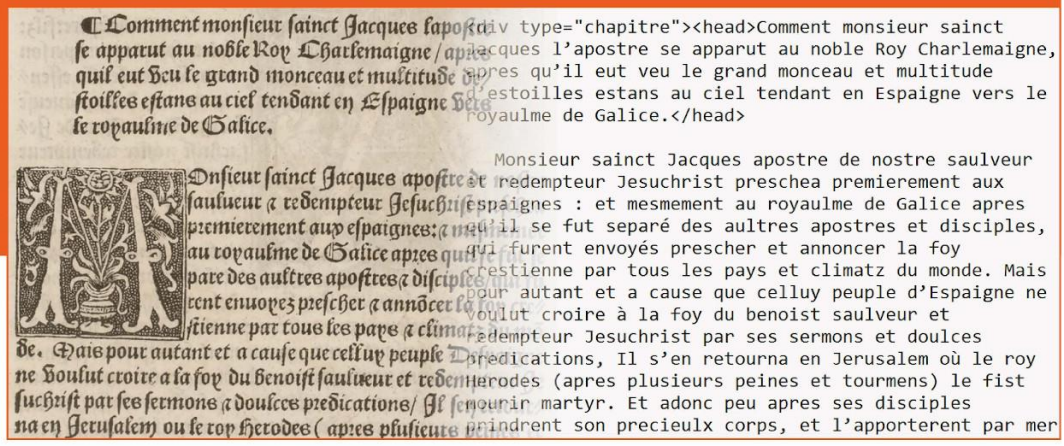
COLLOQUE

13 et 14 octobre 2022

Amphithéâtre MSH-Alpes
Université Grenoble Alpes

La constitution de corpus en diachronie longue

Méthodologies, objectifs et exploitations linguistiques et stylistiques



¶ Comment monſieur ſainct Jacques lapoſtre ſe apparut au noble Roy Charlemaigne / apres quil eut veu le grand monceau et multitude de ſtoilles eſtans au ciel tendant en Eſpaigne vers le royaulme de Galice.

Monſieur ſainct Jacques apoſtre de noſtre ſaulveur et redempteur Jeſuchriſt preſchea premierement aux eſpaignes : et meſmement au royaulme de Galice apres quil ſe fut ſeparé des aultres apoſtres et diſciples, qui furent envoyez preſcher et annoncer la foy cheſtienne par tous les pays et climatz du monde. Mais pour autant et a cauſe que celluy peuple deſpaigne ne voulut croire a la foy du benoiſt ſaulveur et redempteur Jeſuchriſt par ſes ſermons et doulces predications, Il ſen retourna en Jerusalem où le roy Herodes (apres pluſieurs peines et tourmens) le fiſt ſenpourir martyr. Et adonc peu apres ſes diſciples prindrent ſon precieulx corps, et lapporterent par mer

<head>Comment monsieur sainct Jacques l'apostre se apparut au noble Roy Charlemaigne, apres qu'il eut veu le grand monceau et multitude d'estoilles estans au ciel tendant en Espaigne vers le royaulme de Galice.</head>

CONTACTS

julie.sorba@univ-grenoble-alpes.fr // corinne.denoyelle@univ-grenoble-alpes.fr

PROGRAMME

Jeudi 13 octobre 2022

09:00 – 09:30	**ACCUEIL – Petit déjeuner**
09:30 – 10:30	**CONFÉRENCE** *Le corpus SERMO : méthode d'annotation et d'exploitation de corpus paralittéraires pour l'analyse en linguistique diachronique* (Carine Skupien Deken, U. Neuchâtel)
10:30 – 12:30	**SESSION 1 – Constitution de corpus** **(présidence Julie Sorba, UGA, LiDiLEM)**
10:30 – 11:10	*Enjeux des corpus multilingues en diachronie longue : l'exemple du projet MICLE* (Mathieu Goux, UniCaen, CRISCO)
11:10 – 11:50	*Interroger la modalité en latin et en français : construction, annotation et exploitation de corpus* (Cyrielle Montrichard, Helena Bermúdez-Sabel, Corine Rossari & Francesca Dell'Oro, U. Neuchâtel)
11:50 – 12:30	*Corpus rationalisé et dynamique de prose originale en moyen français et français préclassique* (Jean-Michel Jézéquiel, STIH, ATILF)

12:30 – 13:30	**Déjeuner**

13:30 – 14:30	**CONFÉRENCE** *Des corpus à la mesure de la variation du français* (France Martineau, U. Ottawa, Canada)
14:30 – 18:00	**SESSION 2 – Constitution de corpus** **(présidence Pascale Mounier, UGA, Litt&Arts) & Olivier Kraif, UGA, LiDiLEM)**
14:30 – 15:10	*Tagset adaptation to language changing over time. The case of the Electronic Corpus of the 17th- and 18th-century Polish Texts* (Aleksandra Wieczorek, Institute of Polish Language, Polish Academy of Sciences)
15:10 – 15:40	*Enrichissement d'un corpus multiséculaire de théâtre* (Aaron Boussidan, Philippe Gambette, Adrien Roumégous (U. Gustave Eiffel, Laboratoire d'Informatique Gaspard Monge)
15:40 – 16:00	**Pause café**
16:00 – 16:40	*L'évolution de la terminologie artistique à travers l'analyse des traductions en diachronie : la constitution du corpus parallèle plurilingue des Vite de Giorgio Vasari* (Valeria Zotti, U. de Bologne & Daniel Henkel, U. Paris 8, Transferts Critiques anglophones)
16:40 – 17:20	*Antéposition stylistique de l'infinitif et du participe dans l'histoire du français* (Pierre Larrivée & Mathieu Goux, UniCaen, CRISCO)
17:20 – 18:00	*L'atlas Dees électronique* (Tobias Scheer, U. Côte d'Azur, BCL)

Vendredi 14 octobre 2022

09:00 – 09:30	**ACCUEIL – Petit déjeuner**
09:30 – 10:30	**CONFÉRENCE** *La BFM 2022 : un corpus pour les recherches diachroniques en français médiéval et au-delà* (Alexey Lavrentev & Céline Guillot Barbance, ENS Lyon & UMR 5317 IRHIM)
10:30 – 12:30	**SESSION 3 – Constitution de corpus** **(présidence Corinne Denoyelle, UGA, Litt&Arts)**
10:30 – 11:10	*Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval* (Sophie Prévost, Lattice - Loïc Grobol, modyco - Mathieu Dehouck, Lattice - Alexey Lavrentev, IHRIM - Serge Heiden, IHRIM)
11:10 – 11:50	*Un exemple de corpus annoté en diachronie longue : le corpus Democrat, enjeux et exploitations* (Julie Glikman, U. de Strasbourg, LILPA ; Frédéric Landragin, Lattice ; Catherine Schnedecker U. de Strasbourg, LILPA ; Amalia Todirascu, U. de Strasbourg, LILPA)
11:50 – 12:30	*De l'utilité de confronter les sorties de plusieurs étiqueteurs morphosyntaxiques* (Adam Renwick, UGA, LiDiLEM)
12:30 – 14:00	**Déjeuner**
14:00 – 16:00	**SESSION 4 – Effectuer des recherches avec les corpus constitués** **(présidence Adam Renwick, UGA, LiDiLEM)**
14:00 – 14:40	*Observations diachroniques dans un corpus de presse avec le Lexicoscope* (Sascha Diwersy, U. Montpellier 3, PRAXILING & Olivier Kraif, UGA, LiDiLEM)
14:40 – 15:20	*A diachronic corpus to study modality in the Latin language : the WoPoss experience step by step* (Francesca Dell'Oro & Helena Bermúdez-Sabel, U. Neuchâtel)
15:20 – 16:00	*La variation terminologique musicale : Harmonie et polyphonie à la loupe de la textométrie* (Eleonora MARZI, U. de Bologna - Alma Mater Studiorum)
16:00 – 16:15	Mot de clôture (Corinne Denoyelle & Julie Sorba, UGA)

CONFÉRENCE C1
Le corpus SERMO : méthode d'annotation et d'exploitation de corpus paralittéraires pour l'analyse en linguistique diachronique

Carine SKUPIEN DEKENS
U. Neuchâtel - Institut de langue et civilisation françaises, Suisse
carine.skupien-dekens@unine.ch

Dans cette conférence, je présenterai le corpus SERMO, qui contient 62 sermons protestants francophones, édités pour la plupart à Genève, entre 1550 et 1750, représentant env. 600 000 tokens. Après avoir évoqué l'intérêt des sermons pour l'histoire de la langue et de la culture, je présenterai les différents choix méthodologiques qui ont été opérés dans la constitution du corpus et dans son annotation (niveau lexical, morphologique, orthographique, syntaxique et discursif), puis, je montrerai ce qu'on peut y trouver, par différentes méthodes de recherche. Deux aspects seront illustrés plus en détails : le rapport de la langue avec l'oralité, particulièrement importante pour ce genre discursif (Goeury 2018 ; Koch et Oesterreicher 2001 : 586 ; Skupien Dekens 2014, 2018 a et b), et le rapport des prédicateurs avec leur propre discours et avec leur auditoire (Skupien-Dekens 2019, Gerstenberg & Skupien-Dekens 2021), au travers de l'analyse des commentaires métadiscursifs (Hyland 2017) qui ont été balisés systématiquement.

Des exemples de recherches seront réalisés directement sur http://sermo.unine.ch/SERMO/ pour montrer la manière d'obtenir des résultats fiables et adaptés aux différents besoins des chercheuses et chercheurs qui peuvent s'y intéresser.

Références

Gerstenberg, Annette & Skupien Dekens, C. (2021) A Grammar of Authority? – Directive Speech Acts and Terms of Address in Two Single-genre Corpora of Classical French. *Journal of Historical Pragmatics* 22(1). 1–33. 10.1075/jhp.17006.ger.

Goeury, Julien. 2018. Des sermons prononcés comme ils ont été écrits, ou bien écrits comme ils ont été prononcés ?, in *L'Éloquence de la chaire entre écriture et oralité (XIII^e^-XVII^e^ siècle)*, Cinthia Meli (éd.), Paris, Honoré Champion.

Hyland, K. (2017). Metadiscourse: What is it and where is it going? *Journal of Pragmatics* 113, 16-29.

Skupien Dekens C. (2014). Reste-t-il des marques de l'oral dans les sermons de Calvin? In : F. DIÉMOZ *et al. Toujours langue varie... Mélanges de linguistique historique du français et de dialectologie galloromane offerts à M. le Professeur Andres Kristol par ses collègues et anciens élèves*. Genève : Droz, 83-97.

Skupien Dekens C. (2018a). Un genre sous-exploité en histoire du français préclassique et classique, le sermon. In : W. AYRES-BENNETT *et al.* (éds.) *Nouvelles voies d'accès au changement linguistique*. Paris, p.69-84.

Skupien Dekens C. (2018b). La "langue de Canaan" à l'épreuve des sermons (1600-1750). L'exemple des psaumes. *Revue Bossuet*, n° 9, p. 35-58.

Skupien Dekens, C. (2019). L'art de s'adresser à 'l'homme du commun', principes et méthodes du langage simple chez quelques traducteurs et prédicateurs réformés français au XVI^e^ siècle. *Seizième Siècle*, n° 15, p. 121-135.

COMMUNICATION C2
Enjeux des corpus multilingues en diachronie longue : l'exemple du projet MICLE

Mathieu Goux
UniCaen, CRISCO

Les « Très Grands Corpus » se sont multipliés ces dernières années, et leur outillage a permis d'accéder à des données inédites sur les langues et les phénomènes de variation, en français par exemple (Abeillé *et al.*, 2021, Marchello-Nizia *et al.*, 2020). Néanmoins, ces corpus s'appuient généralement sur des données littéraires et se consacrent à une langue donnée, même si parfois au long de plusieurs siècles d'évolution diachronique, ce qui limite nécessairement les résultats accessibles (Gries & Hilpert, 2008, Prévost, 2015, Larrivée & Goux, 2021). Il est cependant possible d'exploiter les avancées de la linguistique de corpus et du TAL pour conduire des recherches dans des corpus multilingues, qu'ils soient constitués de langues génétiquement proches ou issues de familles distinctes, pour faire progresser nos connaissances en grammaire générale.

En prenant l'exemple de l'ANR-DFG MICLE1, qui se consacre à l'évolution de la structure V2 dans les langues romanes du 13e au 17e siècle via l'exemple de l'anglo-normand, de l'ancien et du moyen français, et de l'italien vénitien (Poletto, 2020, Wolfe, 2020), nous souhaitons interroger les enjeux de ces corpus plurilingues pour la recherche en linguistique historique.

Notamment, notre contribution souhaite revenir sur :

- les enjeux d'étiquetages, tant morpho-syntaxiques en parties du discours, que syntaxiques concernant l'analyse de la « phrase », ou de la proposition. Comment évaluer la pertinence des jeux d'étiquettes au regard de ces corpus plurilingues ?
- la chaîne de traitement permettant d'aller du manuscrit à la base de données (Pica, à paraître). Quels logiciels d'analyse peut-on exploiter, quelles précautions méthodologiques faut-il observer et quelles informations pertinentes doit-on conserver pour traiter un grand nombre de données et assurer leur qualité et leur exploitation linguistique ?
- l'exploitation des données et leur visualisation, corrélativement aux formats choisis pour l'encodage et aux logiciels accessibles à la communauté universitaire internationale. Comment assurer la collaboration scientifique et le partage des données ?

En ce sens, notre perspective sera à la croisée des problématiques de recherche et de l'ingénierie. Les deux enjeux sont intimement reliés : les contraintes des bases de données, leur préparation et leur enrichissement déterminent les résultats accessibles, alors que les questions de recherche et la formulation des requêtes influencent en retour la façon dont le corpus sera préparé (Reppen, 2010). Notre contribution sera l'occasion de faire un bilan des possibilités offertes aujourd'hui pour la constitution de corpus multilingues, les avantages et les limites de ces approches, et les enjeux pour la recherche fondamentale en diachronie en prenant l'exemple de la recherche contemporaine en romanistique.

Références

Abeillé, A., Godard, D., Delaveau A. & Gautier, A. (2021). *La grande grammaire du français*. Arles : Actes Sud. https://www.unicaen.fr/projet_de_recherche/micle/ (consulté le 25 mars 2022).

Gabay, S. Clérice, T., Camps, J.-B., Tanguy, J.-B., Gille-Levenson, M. (2020), Standardizing linguistic data: method and tools for annotating (pre orthographic) French, *DDH '20*, 15 17/10 2020, Hammamet.

Galleron, I. & F. Idmhand. (2020). De l'interopérabilité à la réutilisabilité des éditions électroniques. *Humanités numériques* n°1. http://journals.openedition.org/revuehn/ 350

Goux, M. (2021). Text transcription and artificial intelligence: issues & challenges. *Webinar-Seminari di Linguistica : Linguistics Connections*, Universita degli studi di Padova, Dec 2021, Padoue, Italy. https://hal.archives-ouvertes.fr/hal-03500003

Gries, S. Th. & Hilpert, M. (2018). The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, n°3. p. 59-81.

Larrivée, P. & Goux, M. (dir.) (2021). Corpus ConDÉ, version Bêta 1.0, Caen, CRISCO (EA 4255) et PDN (MRSH) de l'Université de Caen. https://www.unicaen.fr/ coutumiers/conde/accueil.html

Manning, D. (2015). Computational Linguistics and Deep Learning. *Computational Linguistics*, n° 41/4. p. 701-707.

Marchello-Nizia, C., Combettes, B., Scheer, T. & Prévost, S. (2020). *Grande grammaire historique du français*. Berlin : De Gruyter.

Pica, M. (à paraître). Harmoniser le corpus ConDÉ : de l'image à la ressource linguistique. *Studia Linguistica Romanica*, n° spécial « Le temps long : l'évolution du français dans un corpus textuel calibré. Le témoignage de la Coutume de la Normandie (1250-1771) ».

Poletto, C. (2020). More than one way out. On the factors influencing the loss of V to C movement. *Linguistic Variation* n°19/1. p. 47-81.

Prévost, S. (2015). Diachronie du français et linguistique de corpus : une approche quantitative renouvelée. *Langages*, n°197. p. 23-45.

Reppen, R. (2010). Building a corpus. What are the key considerations? A. O'Keeffe & M. McCarthy (éd.), *The Routledge Handbook of Corpus Linguistics*. Londres: Routledge. p. 31-37.

Wolfe, S. (2020). Redefining the typology of V2 languages. The view from Medieval Romance and beyond. *Linguistic Variation*, n°19/1. p. 16-46.

COMMUNICATION C3

Interroger la modalité en latin et en français : construction, annotation et exploitation de corpus

Cyrielle MONTRICHARD, Helena BERMUDEZ-SABEL, Corinne ROSSARI & Francesca DELL'ORO
U. Neuchâtel, Suisse

Notre contribution s'inscrit dans un projet visant l'étude de la modalité en latin et en français dans une perspective de linguistique outillée de corpus. La modalité est une notion qui recouvre une grande diversité de formes dont la plupart sont susceptibles de remplir différentes valeurs modales. L'objectif de notre recherche est d'analyser le système modal dans deux langues séparées par un empan diachronique conséquent, mais reliées génétiquement. Cette démarche que nous considérons *achronique* – dans la mesure où il n'est pas question de retracer l'évolution d'une langue à l'autre, mais de comparer des états de langue ponctuels – permettra d'interroger la modalité sous un angle original en relevant d'une part ce qui est propre au système linguistique du latin et du français indépendamment l'un de l'autre ; et d'autre part en tissant des liens entre les deux systèmes au moyen de l'interrogation statistique de données textuelles constituées en corpus.

Pour atteindre cet objectif, la constitution de notre corpus comprend les deux variables suivantes : (i) les types de genre discursif et (ii) deux états de langue propres à chaque système. Ces deux variables permettent de cerner ce qui relève de l'influence du genre sur l'usage de la modalité et de tenir compte des changements propres à chaque langue via les deux tranches chronologiques envisagées.

Cet objectif nous met face à plusieurs défis. Un premier défi se situe dans le choix des textes, puisque nous ambitionnons de construire des corpus comparables du point de vue des genres tant entre les deux langues qu'entre les deux tranches chronologiques de chaque langue. Un autre défi se situe du côté de la conception d'un système d'annotation qui permette d'établir une comparaison sur le plan sémantique. Cette tâche est en effet complexe car différents aspects doivent être considérés :

- les formes de la modalité posent problème quant à leur identification : les formes comme les flexions temporelles, adverbes, verbes… ne sont pas nécessairement équivalentes entre les deux langues ;
- les sens que ces formes sont susceptibles de véhiculer ne se superposent pas non plus : on sait que les notions dégagées traditionnellement, comme la volition, la possibilité, la nécessité, ont des contours flous, en témoigne la variation selon les cadres théoriques (Le Querler, 2001 ; Narrog, 2012 ; Nuyts, 2016).
- la polysémie intrinsèque à plusieurs catégories de formes modales : le nombre et choix des valeurs qu'il est pertinent de retenir pour un verbe comme *devoir* et *debere* ne se recouvrent pas nécessairement.

Ainsi, le projet d'annotation doit (i) concevoir des schémas d'annotation compatibles entre les langues et les états de langue ; (ii) intégrer à ces schémas différents niveaux de granularité pour saisir les particularités propres à chaque système ; (iii) mobiliser des outils communs aux deux langues permettant une annotation automatique, semi-automatique et manuelle pour garantir la comparabilité des deux langues.

Notre contribution commencera par décrire les choix opérés pour relever les défis soulevés, avant de présenter les premiers résultats issus des recherches effectuées au moyen de la plateforme TXM (Heiden, 2010).

Remerciements

Ce travail a bénéficié d'un financement du Fonds Jakob Wüest, à travers la fondation Empiris.

Références

Heiden, S. ; Magué, J.-P. & Pincemin, B. (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, Italie. pp. 1021-1032. http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf

Le Querler, N. (1996). *Typologie des modalités*. Caen : Presses universitaires de Caen.

Narrog, H. (2012). *Modality, subjectivity, and semantic change*. Oxford: Oxford University Press.

Nuyts, J. (2016). Analyses of the Modal Meanings. In J. Nuyts & J. van der Auwera (éds.), *The Oxford Handbook of Modality and Mood*. Oxford: Oxford University Press. 31-49. https://doi.org/10.1093/oxfordhb/9780199591435.013.1.

COMMUNICATION C4
Corpus rationalisé et dynamique de prose originale en moyen français et français préclassique

Jean-Michel JEZEQUEL
STIH, ATILF (UMR 7118)

Cette communication présentera les enjeux méthodologiques qui, sous-tendus par une question de recherche précise, nous ont conduit à l'établissement d'un large corpus représentatif de la langue et échantillonnable à discrétion. Pour notre étude syntaxique, en moyen français et français préclassique, de la variation de la préposition précédant un infinitif objet d'un verbe recteur, trois paramètres supplémentaires sont venus s'ajouter aux exigences inhérentes à la constitution d'un corpus représentatif de la langue : l'éviction de textes versifiés, de textes ayant pour modèle un texte d'une autre langue — les traductions ou réécritures —, et de textes en anglo-normand.

Positivement, le corpus attendu devait être un corpus de prose originale représentative des états de langue du moyen français (1330-1550) et du français préclassique (1550-1630). De surcroît, nous souhaitions un suivi chronologique fin pour essayer de tendre à des statistiques par quart de siècle. Les récents corpus en diachronie longue de la GGHF et du projet Profiterole se sont avérés insuffisants à cause de ces quatre contraintes de sélection supplémentaires, mais leur rationalisation a servi de base méthodique à la construction de notre corpus de prose française originale. En parcourant l'ensemble des bases numériques actuellement à la disposition du diachronicien du français, 41 cotes ont été retenues, principalement issues de Frantext (environ 60%), mais aussi de la BFM, du corpus *Phraséologie du roman médiéval* et du corpus *Récits de voyage* de Capucine Herbert. L'ensemble de ces textes en prose ressortissent à sept sous-genres différents, répartis aussi harmonieusement que possible au cours du temps, plus ou moins tous les dix ans, de 1298 à 1627, de façon à réaliser un « bon "maillage" du corpus » (PREVOST, 2015). Le corpus compte, à l'été 2022, 45 cotes, grâce à quatre textes que nous avons édités nous-même, et devrait s'élever définitivement en octobre 2022 à 49 cotes à la faveur de la numérisation de quatre chroniques encore numériquement inédites.

En matière d'édition des textes, l'exigence n'a pas pu être poussée systématiquement jusqu'à la fidélité graphique. L'ensemble du corpus a été annoté automatiquement en lemme et en étiquette morphosyntaxique Cattex par l'outil LGeRM. Enfin, une fonctionnalité inconnue de Frantext ou de la BFM a été créée dans l'outil LGeRM qui assure l'exploitation du corpus : l'échantillonnage dynamique. Il est ainsi possible d'échantillonner individuellement et à volonté chaque cote du corpus. En évitant la sur-ou sous-représentation d'un texte, d'un genre ou d'une date, l'échantillonnage participe au premier chef à la représentativité de la langue d'un corpus. Et en étant dynamique, l'échantillonnage peut s'adapter à l'objet d'étude et à la méthodologie de chaque recherche. À titre d'exemple, l'échantillonnage dynamique nous a permis de trouver un point d'équilibre à un million de tokens (mots et ponctuations) répartis sur 330 ans pour notre recherche en syntaxe qui, d'une part, demande un corpus de grande taille car les occurrences sont de faible fréquence (VAN REENEN & SCHØSLER, 1993), mais où, d'autre part, chaque occurrence doit être lue, analysée et annotée manuellement. Initié par une question de recherche précise en syntaxe, ce large corpus, potentiellement échantillonnable jusqu'à deux millions de tokens (voire davantage si recours à des coefficients de pondération), pourrait se prêter à d'autres objets d'étude en syntaxe, morpho-syntaxe et lexicologie.

Références

Bases de textes numériques

BFM - Base de Français Médiéval [En ligne]. Lyon : ENS de Lyon, Laboratoire IHRIM, 2019, http://bfm.ens-lyon.fr

Corpus de la *GGHF*, in MARCHELLO-NIZIA, C., COMBETTES, B., PREVOST, S., & SCHEER, T. (éd.) (2020). *Grande grammaire historique du français*. Boston : De Gruyter Mouton, p. 46- 50.

Corpus du projet ANR Profiterole. Paris : CNRS, Laboratoire Lattice, [en cours], https://www.lattice.cnrs.fr/projets/projet-anr-profiterole/

Corpus du projet *Phraséologie du roman médiéval : une approche sur corpus outillée*. Grenoble : Université Grenoble Alpes, Laboratoires LIDILEM et Litt&Arts, [en cours], https://hal.archives-ouvertes.fr/hal-02147699

Frantext – Base textuelle Frantext [En ligne]. Nancy : CNRS, Laboratoire ATILF, 2022, https://www.frantext.fr

Récits de voyage, corpus établi par Capucine Herbert [En ligne]. Nancy : CNRS | Université de Lorraine, Laboratoire ATILF, 2015, http://www.atilf.fr/dmf/RecitsVoyage

Outils

LGeRM - Plateforme de lemmatisation de la variation graphique des états anciens du français [En ligne]. Nancy : CNRS | Université de Lorraine, Laboratoire ATILF, http://www.atilf.fr/LGeRM/

Cattex – Jeu d'étiquettes morpho-syntaxiques pour la langue française médiévale. Lyon : ENS de Lyon, Laboratoire IHRIM, 2009, http://bfm.ens-lyon.fr/spip.php?article176

Articles

PREVOST, S. (2015). Diachronie du français et linguistique de corpus : une approche quantitative renouvelée, *Langages*, vol. 197 (1), p. 8. Lien vers l'article : halshs- 01423562. V

AN REENEN, P., & SCHØSLER, L. (1993). Les indices d'infinitif complément d'objet en ancien français, in Lorenzo, R. (éd.) *Actas do XIX Congreso Internacional de Lingüística e Filoloxía Románicas, Universidade de Santiago de Compostela, 1989. V, Gramática histórica e historia da lingua*. A Coruña: Fundación « Pedro Barrié de la Maza, Conde de Fenosa, » p. 524-525.

Corpus rationalisé et dynamique de prose originale en moyen français et français préclassique

	source	cote	date	valid° date	échant	genre	début échant	fin échant	cote entière
		45	330 ans		928000				4 757 044
▼ 13e		1			50				
	FRTX	ActesFerry3	1298	x	50	textes de la pratique	123008	173008	178936
▼ 14e		13			282				
	GRE	ArtusBretagne	1300		20	récit littéraire		20000	215797
	GGHF	Joinville	1305		20	chronique	17052	37052	84175
	FRTX	ChrMorée	1322		20	chronique		20000	106120
	FRTX modif	Magloire1330-1341	1335	x	28	textes de la pratique			27983
	JMJ	GChrPh6fin 1344-1350	1350		20	chronique			20499
	FRTX	Berinus	1360		20	récit littéraire		20000	119126
	FRTX modif	MachautVoirDitLettres	1364		20	correspondance		20000	33083
	FRTX	OresmeComEthique	1370		30	texte d'idée		30000	123521
	BFM	OrgemontChV	1381?	?	14	chronique			14219
	FRTX	PhebusChasse	1389		20	savoir de spécialité	34685	54685	86695
	FRTX	RegChâtelet	1389	x	30	textes de la pratique		30000	200047
	FRTX	FroissartL3	1390		20	chronique		20000	74055
	FRTX	Mélusine	1392		20	récit littéraire		20000	141476
▼ 15e		14			278				
	FRTX	XVjoies	1400		20	récit littéraire		20000	39530
	FRTX	BayeGreffier	1403	x	20	textes de la pratique	4742	24742	87647
	BFM	CoutForêts	1409		30	textes de la pratique		30000	157976
	FRTX	PizanPaix	1412		20	texte d'idée	1243	21243	61843
	FRTX	FauquembergueGreffier	1418	x	20	textes de la pratique	11329	31329	85765
	HERB	PèlerinageSinaï	1425		13	récit de voyage			12652
	HERB	PèlerinageVelaines	1432		15	récit de voyage			15022
	BFM	MonstreletChr	1444	?	20	chronique		20000	32747
	FRTX	Saintré	1456		20	récit littéraire		20000	105425
	FRTX	Jouvencel	1461		20	récit littéraire		20000	58530
	FRTX	RoyeChrScandaleuse	1465	x	20	chronique	7077	27077	58530
	BFM	CNN Anonyme ?	1467		20	récit littéraire	6975	26975	176741
	HERB	Barbartre	1480		20	récit de voyage	1200	21200	35252
	FRTX	Commynes1	1489		20	chronique		20000	71364
▼ 16e		10			181				
	FRTX	Bouchart	1514		20	chronique	55000	75000	157695
	FRTX FALT.	VigneullesMémoires	1522		20	chronique	21045	41045	190866
	FRTX	Crignon	1529		19	récit de voyage			18884
	FRTX	CalvinLettres	1544	x	10	correspondance		10000	37537
	FRTX	BelonOyseaux	1555		20	savoir de spécialité		20000	185858
	FRTX	Navarre	1559		20	récit littéraire		20000	205325
	FRTX modif	Pasquier1554-1567	1559	x	12	correspondance			12064
	FRTX	Léry	1580		20	récit de voyage	73000	93000	129210
	JMJ	ParéL19	1585		20	savoir de spécialité		20000	28955
	FRTX	MontaigneEssais	1592		20	texte d'idée			343799
▼ 17e		7			137				
	JMJ	Champlain	1603		17	récit de voyage			17028
	JMJ	Fourier	1606	x	10	correspondance	2360	12360	157215
	FRTX	UrféAstrée	1612		30	récit littéraire		30000	242549
	FRTX	GuezBalzac	1622	x	10	correspondance		10000	56004
	FRTX	CamusPalombe	1625		30	récit littéraire		30000	131167
	FRTX	SorelBerger	1627		30	récit littéraire		30000	325509
	FRTX	PeirescLettresDupuy	1627	x	10	correspondance	31870	41870	92623

x : un calcul de moyenne (ou un choix de date) a été opéré
gris : la moyenne peut varier selon l'empan de l'échantillon

Répartition des échantillons par genre et par siècle

siècle	13e	14e	15e	16e	17e	Total général
genre	928000 (Somme)					
chronique		94	60	40		194
correspondance		20		22	30	72
récit de voyage			48	39	17	104
récit littéraire		60	80	20	90	250
savoir de spécialité		20		40		60
texte d'idée		30	20	20		70
textes de la pratique	50	58	70			178
Total général	50	282	278	181	137	928

Nb de cotes par source

source	source (Tout compter)	source (Tout compter)
FRTX	28	62 %
BFM	4	9 %
JMJ	4	9 %
FRTX modif	3	7 %
HERB	3	7 %
FRTX FALT.	1	2 %
GGHF	1	2 %
GRE	1	2 %
Total général	45	100 %

Nb de cotes par genre

genre	genre (Tout compter)
chronique	10
correspondance	6
récit de voyage	6
récit littéraire	11
savoir de spécialité	3
texte d'idée	3
textes de la pratique	6
Total général	45

CONFÉRENCE C5
Des corpus à la mesure de la variation du français

France MARTINEAU
U. Ottawa, Canada

Ma conférence portera sur la constitution de trois grands corpus historiques que j'ai établi avec des équipes de recherche, le *Corpus MCVF* (IX^{e} s.-XVII^{e} s.) (Martineau *et al*., 2010) le *Corpus FRAN* (XVIII^{e} s.-XXI^{e} s.) (Martineau, *et al*. 2011 ; Martineau et Séguin, 2016) et le *Corpus LFFA* (XVII^{e} s.-XXI^{e} s.) (Martineau, 1995, 2005, 2012).

Je discuterai dans un premier temps de l'articulation entre ces trois corpus en mettant l'accent sur leur complémentarité pour situer la variation diachronique, diatopique et diastratique dans l'espace et le temps. Je m'attarderai notamment aux questions suivantes : représentativité et périodisation, métadonnées, arrimage de corpus textuels et oraux et pérennisation. J'illustrerai l'intérêt de tels corpus à partir du changement et de la variation dans la négation (négation simple ou bipartite/concordance négative/effacement de *ne*, par ex.) et présenterai brièvement l'ouvrage *Grammaire et lexique historiques du français nord-américain* (2022) en situant cet ouvrage dans la diachronie longue du français.

Références

Martineau, France (1995), *Corpus du Laboratoire de français familier ancien*, Université d'Ottawa.

Martineau, France (2005), Perspectives sur le changement linguistique : aux sources du français, *Revue canadienne de linguistique*, vol. 50, n^{os} 1-4, p. 173-213.

Martineau, France *et al*. (2010), *Corpus MCVF (Modéliser le changement : les voies du français)*, Université d'Ottawa.

Martineau, France *et al*. (2011), *Corpus FRAN* (Français d'Amérique du Nord), Université d'Ottawa/Université de Sherbrooke.

Martineau, France (2012), Les voix silencieuses de la sociolinguistique historique, *Cahiers de linguistique*, vol. 38, n^{o} 1, p. 111-135.

Martineau, France, et Marie-Claude Séguin (2016), Le *Corpus FRAN* : réseaux et maillages en Amérique française, *Corpus*, vol. 15, p. 55-87.

SESSION 2 – Constitution de corpus

COMMUNICATION C6

Tagset adaptation to language changing over time. The case of the Electronic Corpus of the 17th- and 18th-century Polish Texts

Aleksandra WIECZOREK
Institute of Polish Language, Polish Academy of Sciences, Pologne

The Electronic Corpus of the 17th- and 18th-century Polish Texts is a richly annotated corpus of Middle Polish containing around 13.5M tokens (in the near futur 25M; www.korba.edu.pl). The corpus covers the Baroque era as well as a large part of the Enlightenment. When creating such a corpus that covers a long period, the changes that were taking place in the language in those years should be considered. In the era of our interest, some new grammatical categories or their values have appeared in Polish, and others have disappeared. This development of the grammatical system and language differentiation over time should be reflected in the morphosyntactic annotation of the corpus – in the structure of the tagset, as well as in the use of individual tags. The team's experience has shown that there is no universal solution to this problem and that for each grammatical feature, different issues should be taken into account. These include substantive issues, such as the degree of prevalence of a given linguistic phenomenon (whether it concerns a few forms or is it a serial phenomenon), the timeframe for linguistic change, as well as practical issues, such as the possibility of creating useful corpus queries and conducting linguistic research in future.

The first example of grammatical challenges that our team encountered is the variation of some inflectional endings of masculine animate nouns in the Polish language of the period under discussion. It was the time of formation of a new grammatical category, called "masculine-personality" (pl. męskoosobowość). In modern Polish, masculine nouns in nominative, accusative and vocative plural have different grammatical endings depending (simplifying a bit) on the meaning – masculine personal nouns take one type of endings, while the rest masculine nouns take another type of endings. On the contrary, in Old Polish (before 1500) the grammatical endings of N, A and V pl forms of masculine nouns were not related to semantics at all. In between, in the Middle Polish, we observe a vast variation of endings in N, A and V pl forms of animate masculine nouns (both personal and non-personal) – sometimes they were declined according to the older patterns, and sometimes according to modern ones. Therefore the tagset used to annotate the corpus, which was originally created for the corpus of modern Polish, had to be slightly changed in this regard.

The second example of a grammatical difficulty is the disappearance of the dual number in the 18th-century Polish language. At the same time, some forms with the former dual ending survived, but became plural forms. Of course, such linguistic changes do not happen overnight or even within a year. It may be assumed that at the beginning of our epoch forms with former dual endings were treated as forms of dual number, at the end – as forms of plural, but we cannot be sure of their interpretation in the texts that arose between these time points. Here, an arbitrary decision was made to end marking dual number in texts published after 1740.

Références

Bronikowska, R., Gruszczyński, W., Ogrodniczuk, M., Woliński, M. (2016). *The Use of Electronic Historical Dictionary Data in Corpus Design*. *Studies in Polish Linguistics* 11:47–56.

Electronic Corpus of the 17th and 18th c. Polish Texts (up to 1772). http://korba.edu.pl

Gruszczyński, W., Adamiec, D., Bronikowska, R., Kieraś, W., Modrzejewski, E., Wieczorek, A., Woliński, M. (2022). *The Electronic Corpus of 17th- and 18th-century Polish Texts*, "Language Resources and Evaluation" 56, pp. 309–332.

Kieraś, W., Komosińska, D., Modrzejewski, E., Woliński, M. (2017). *Morphosyntactic Annotation of Historical Texts. The Making of the Baroque Corpus of Polish*. In: K. Ekštein, V. Matoušek (eds.) 20th International Conference on Text, Speech, and Dialogue (TSD 2017). Lecture Notes in Computer Science 10415, pp. 308–316. Springer, Cham.

Kucała, M. (1978). *Rodzaj gramatyczny w historii polszczyzny*, Wrocław – Zakład Narodowy im. Ossolińskich, Wydawnictwo Polskiej Akademii Nauk.

Rzepka, W.R. (1975). *Dopełniacz w funkcji biernika męskich form osobowych w liczbie mnogiej w polszczyźnie XVII w.* Wrocław – Zakład Narodowy im. Ossolińskich, Wydawnictwo Polskiej Akademii Nauk.

Rzepka, W.R. (1985). *Demorfologizacja rodzaju w liczbie mnogiej rzeczowników w polszczyźnie XVI-XVII wieku*. Poznań – Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza.

COMMUNICATION C7
Enrichissement d'un corpus multiséculaire de théâtre

Aaron BOUSSIDAN, Philippe GAMBETTE & Adrien ROUMEGOUS
LIGM (UMR 8049), U. Gustave Eiffel.

Le corpus réuni par Paul Fièvre au format XML-TEI sur le site theatre-classique.fr a été utilisé dans plusieurs études en humanités numériques ou en traitement automatique des langues (Karsdorp *et al.* 2015 ; Douguet, 2018), parfois après un reformatage pour le rendre plus standard et lui appliquer divers outils de visualisation (Glorieux 2016 ; Fischer *et al.* 2019). Nous nous intéressons à l'enrichissement de ce corpus à partir d'autres collections de pièces de théâtre constituées dans le cadre de travaux étudiants (Canu & Carpentier 2021), de projets de recherche comme le corpus Malherbe du projet Anamètre (Renault 2018), de relectures collaboratives comme sur Wikisource ou individuelles comme celles de Michel Capus sur le site théâtre-documentation.com.

Cet enrichissement présente à la fois des enjeux quantitatifs et des enjeux qualitatifs de représentativité. Nous évaluons ces enjeux sur la base du corpus constitué par Céline Fournial pour étudier de façon systématique les sources du théâtre classique français (Fournial 2019). Hyperpièces, disponible sur celinefournial.github.io/hyperpieces, est composé de 104 comédies, 216 tragi-comédies et 273 tragédies, sur la période 1550-1650. Le corpus French DraCor, dérivé de celui de theatre-classique.fr couvre environ un quart de ce corpus, avec une surreprésentation des comédies : près d'un tiers des comédies référencées dans Hyperpièces sont contenues dans DraCor. Quant aux corpus de théâtre-documentation (TD), Bibliothèque dramatique (BD) et Wikisource (WS), ils couvrent respectivement environ 17%, 10% et 1% du corpus Hyperpièces. Par ailleurs, les corpus TD et BD apportent chacun une vingtaine de pièces absentes de DraCor. Le corpus WS a davantage d'intérêt sur les siècles plus récents.

Se posent alors diverses questions, liées à des problématiques techniques (détection des doublons entre les différents corpus et conversions de format nécessaires pour aboutir à celui du corpus de référence) ou à l'uniformité éditoriale du corpus, par exemple pour les didascalies (Galleron 2021). En particulier, pour les pièces des siècles les plus anciens, la question de la normalisation de la langue a été gérée différemment selon les sources.

Nous proposons une chaine de traitement informatique visant à uniformiser le format des pièces de théâtre collectées, en adoptant les choix effectués au sein du projet DraCor et en extrayant des données sources les informations requises pour compléter les métadonnées et structurer correctement le texte des pièces. Elle est constituée par un ensemble de scripts Python mis à disposition sous licence libre sur github.com/AaronFive/StageHyperpieces. Un outil de normalisation automatique fondé sur des règles (Bawden *et al.* 2022) nous permet d'estimer l'état de langue des pièces ajoutées et en proposer une caractérisation, en quantifiant le nombre de règles appliquées.

Nous montrons enfin comment ces questions de normalisation peuvent être contournées en utilisant ces corpus pour des analyses qui s'intéressent à la structure des pièces plutôt qu'à leur contenu textuel. Plusieurs objets mathématiques peuvent en effet être utilisés pour représenter la structure d'une pièce de théâtre, notamment des réseaux de personnages (Lotker 2021), matrices de co-présences (Marcus 1970, Brainer & Neufeldt 1974, Douguet 2015) ou encore modèles de mots paramétrés (Boussidan 2021).

Références

Bawden, Rachel ; Poinhos, Jonathan ; Kogkitsidou, Eleni ; Gambette, Philippe ; Sagot, Benoît ; Gabay, Simon. Automatic Normalisation of Early Modern French. *LREC 2022 - 13th Language Resources and Evaluation Conference*, 2022.

Brainerd, Barron ; Neufeldt, Victoria. On Marcus' methods for the analysis of the strategy of a play. *Poetics*, 3(2) :31–74, 1974.

Boussidan, Aaron. *Modélisation de pièces de théâtre*, mémoire de master 2, Université Gustave Eiffel, 2021.

Canu, Amélie ; Carpentier, Claire. La Bibliothèque dramatique. L'édition numérique d'un corpus de pièces de théâtre du XVII[e] siècle. *Dix ans avec CAHIER : des corpus d'auteurs pour les humanités à leur exploitation numérique*, 2021.

Douguet, Marc. *La composition dramatique : La liaison des scènes dans le théâtre français du XVII[e] siècle,* thèse de doctorat en Langues et littératures françaises, Université Paris 8, 2015.

Douguet, Marc. Les hémistiches répétés. *JADT'18* : 215, 2018.

Fischer, Frank ; Börner, Ingo ; Göbel, Mathias ; Hechtl, Angelika ; Kittel, Christopher ; Milling, Carsten ; Trilcke, Peer. Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. *Proceedings of DH2019*.

Fournial, Céline. *Imitation et création dans le « théâtre moderne » (1550-1650) : la question des cycles d'inspiration,* thèse de doctorat en littérature et civilisation française, Sorbonne Université, 2019.

Galleron, Ioana. Pour un balisage sémantique des textes de théâtre : le cas des didascalies. *Sens public*, 2021.

Glorieux, Frédéric. Dramagraphie 0.2. Carnet Hypothèses *J'attends des résultats*.

Karsdorp, Folgert ; Kestemont, Mike ; Schöch, Christof ; van den Bosch, Antal. The love equation: Computational modeling of romantic relationships in French classical drama. *6th Workshop on Computational Models of Narrative (CMN'15)*, 2015.

Lotker, Zvi. *Analyzing Narratives in Social Networks*, Springer Cham, 2021.

Marcus, Solomon. *Mathematical poetics*. Bucharest, The Publishing House of the SRR Academy (1970).

Renault, Richard. Corpus Malherbe : corpus de textes versifiés du XVIIe au début du XXe. *Journée d'étude CORLI : Traitements et standardisation des corpus multimodaux et web 2.0*, 2018.

COMMUNICATION C8
L'évolution de la terminologie artistique à travers l'analyse des traductions en diachronie : la constitution du corpus parallèle plurilingue des Vite de Giorgio Vasari

Valeria ZOTTI
Università di Bologna, Italie
Daniel HENKEL
U. Paris 8, TransCrit

Nous présentons le projet de constitution d'un corpus parallèle des traductions d'un texte italien de référence de la Renaissance pour la description du patrimoine artistique florentin : *Les Vies des meilleurs peintres, sculpteurs et architectes* de Giorgio Vasari. Cet ouvrage est considéré comme le premier ouvrage critique d'art de l'histoire, dans lequel les artistes choisis par l'auteur, artiste célèbre à son tour, passent à la postérité (Vasselin, *Encyclopædia Universalis*). Les deux « âmes » de la terminologie artistique, humaniste et technique, coexistent dans ce texte fondateur de Vasari qui est, à juste titre, reconnu comme l'initiateur de la langue de la critique d'art italienne (Biffi 2017 : 28).

Dans le cadre du projet de recherche interuniversitaire *Lessico Plurilingue dei Beni Culturali* (lessicobeniculturali.net), mené à l'Université de Florence en collaboration avec d'autres Universités italiennes et étrangères, dont l'objectif est la création d'un dictionnaire plurilingue du lexique du patrimoine culturel, nous sommes chargés de la constitution des corpus parallèles des différentes traductions (allemand, anglais, français, espagnol, russe, avec l'italien comme langue de départ), qui se sont succédé au cours des siècles (en particulier du XVIII^{e} au XX^{e} siècle), de cet ouvrage qui a contribué à la naissance d'un lexique artistique européen. Un corpus comparable multilingue a été déjà réalisé au sein de ce projet et est disponible sur http://corpora.lessicobeniculturali.net (Billero *et al.* 2020).

Le corpus parallèle italien-français, sur lequel nous nous arrêtons plus en détail, sera composé notamment de l'une des deux éditions italiennes des *Vies* de Vasari (1550 et 1568), et de quatre traductions françaises (Leclanché/Jeanron 1839, Weiss 1900, Chastel 1981, Luciani 2002), deux autres n'étant pas exploitables pour différentes raisons. Chaque traduction atteste le développement et aussi la diversification, selon le public cible, du langage de la critique d'art en France dans les différents stades de son évolution : de l'intégration dans un lexique professionnel rudimentaire de nombreux emprunts à l'italien dans le XVII^{e} siècle jusqu'à la stabilisation d'une terminologie cultivée quelques siècles plus tard (Chastel 1981) et à sa vulgarisation (Luciani 2022).

Dans cette communication nous nous penchons sur la méthodologie et sur les critères adoptés pour la constitution des bases de données parallèles italien-anglais et italien-français, en cours de réalisation : à partir de la préparation des fichiers à employer (pré-alignement), en passant par l'alignement automatique à l'aide du logiciel LF Aligner, puis la correction manuelle des tableaux d'alignement afin d'établir une correspondance parfaite entre texte source et texte cible nécessitant des règles de segmentation différentes pour les deux, la conversion du tableau corrigé au format .tmx (Translation Memory Exchange) et, pour finir, à une deuxième vérification semi-automatique dans Okapi Checkmate, afin de repérer tout écart de longueur suspect entre segments source et cible.

Nous illustrerons en outre, par quelques exemples, le type de recherche linguistique que permettra dans le futur ce corpus parallèle en diachronie pour mettre en lumière l'évolution de la terminologie artistique en français au cours des siècles et, cela, à travers des modes d'interrogation spécifiques offerts par un logiciel, comme HyperMachiavel (Gedzelman et Zancarini 2011), qui peut être exploité non seulement en tant qu'outil d'alignement de segments textuels et de comparaison des traductions,

tant en synchronie qu'en diachronie, mais aussi pour la recherche, l'étiquetage semi-automatique et l'analyse conceptuelle des équivalents traductionnels.

Références

Biffi, Marco (2017), *L'italiano dell'arte*, in *L'Italiano. Conoscere una lingua formidabile*, Accademia della Crusca, Roma, GEDI.

Billero, Riccardo; Farina, Annick; Nicolás Martínez, Maria Carlota (eds.) (2020), *I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali*, Firenze University Press, Firenze, coll. Strumenti per la didattica e la ricerca, Italie.

Gedzelman, Séverine ; Zancarini, Jean-Claude (2011), *HyperMachiavel : un outil de comparaison de traductions*, « Lingua e stile », vol. XLVI, No 2, Società editrice Il Mulino, Italie, p. 247-266. doi : 10.1417/36054

Zotti, Valeria (2017), *L'integrazione di corpora paralleli di traduzione alla descrizione lessicogra_ca della lingua dell'arte : l'esempio delle traduzioni francesi delle Vite di Vasari*, in Zotti, Valeria ; Pano Alaman, Ana (dir.) Informatica umanistica. Risorse e strumenti per lo studio del lessico dei beni culturali, Firenze University Press, coll. Strumenti per la didattica e la ricerca, Italie, p. 105-134.

Corpus parallèle français-italien

Vasari, Giorgio, *Le vite de' piu eccellenti architetti, pittori, et scultori italiani, da Cimabue insino a'tempi nostri*, Lorenzo Torrentino, Italia, 1550, 994 p.

Vasari, Giorgio, *Le vite de' piu eccelenti pittori, scultori ed architettori*, Giunti, Italia, 1568, 523 p.

Vasari, Giorgio, *Les vies des meilleurs peintres, sculpteurs et architectes*. Traduction française et édition commentée sous la direction d'André Chastel, Berger-Levrault, coll. Arts, France, 1981, 12 volumes.

Vasari, Giorgio, *Vies des peintres, sculpteurs et architectes par Giorgio Vasari* : traduites par L. Leclanché et commentées par P.-A. Jeanron et L. Leclanché., Tessier, France, 1839-1842, 10 tomes.

Vasari, Giorgio, *Les vies des plus excellents peintres, sculpteurs, et architectes*. Traduction par Charles Weiss, Dorbon-Aine, France, 1900, 3 volumes.

Vasari, Giorgio ; Benvenuto, Cellini, *Vies d'artistes*. Édition et traduction de l'italien par Gérard Luciani, Gallimard, coll. Folio bilingue 105, France, 2002, 432 p.

COMMUNICATION C9
Antéposition stylistique de l'infinitif et du participe dans l'histoire du français

Pierre LARRIVÉE & Mathieu GOUX
Unicaen, CRISCO

Le but de cette présentation est d'illustrer l'usage d'un corpus doublement calibré pour suivre l'évolution d'une construction syntaxique pour 4 siècles de l'histoire du français. Le premier sous-corpus est un ensemble de pièces dialogales de quatre procès en prose de la région (anglo-)normande de 1275, 1340, 1430 à 1591. Le second relève du roman courtois en prose avec des textes de 1250, 1350, 1461 et 1550. Ce choix permet de vérifier des différences de construction et de fréquence sous l'angle du registre à travers le temps, étant entendu que le second ensemble, littéraire, est probablement d'un registre plus soutenu que les dialogues légaux.

La construction étudiée est celle de l'infinitif et du participe antéposé au verbe modal conjugué, type *Autant que faire se peut*, connu généralement sous le nom de Fronting ou de Déplacement stylistique. S'appuyant sur une annotation syntaxique en dépendance faite via le l'analyseur algorithmique HOPS et corrigée manuellement, les occurrences sont extraites et analysées quant à leur distribution et aux éléments récurrents du contexte, en contraste avec les infinitifs et participe postposés aux modaux.

L'analyse établit les résultats suivants : une fréquence très faible, et plus faible dans les textes non littéraires ; une distribution essentiellement dans les propositions à sujet nul en subordonnée dans les deux sous-corpus ; une perte graduelle du rendement fonctionnel de la construction qui devient essentiellement représentée par le verbe vicariant *faire* précédant *pouvoir*. On confirme en outre la condition d'ordre des mots harmonique observée pour l'italien observée par Poletto et Pinzin (en cours) : l'objet de l'infinitif doit précéder l'infinitif si celui-ci précède le modal, excluant la séquence [Infinitif] [Objet] [Modal]. Enfin, concernant la question de savoir ce qui motive l'antéposition elle-même, l'idée qu'elle marquerait une valeur informationnelle (verum focus du verbe, Rahn 2016, Dufter 2018) est explorée. Le travail montre qu'un corpus calibré permet de mieux suivre la courbe d'évolution, et qu'un corpus doublement calibré permet par comparaison de s'approcher de la compétence immédiate des locuteurs à époque ancienne.

Références

De Andrade, Aroldo. 2018. Aboutness Topics in Old and Middle French: A corpus-based study on the fate of V2. *Canadian Journal of Linguistics/Revue canadienne* 63,2, 194-220.

Combettes Bernard. 2006. L'analyse thème/rhème dans une perspective diachronique. *LINX* 55, 75-90.

Dufter, Andreas. 2018. *On participle fronting in Old and Middle French*. Ppt, Universität München.

Fischer, S. et A. Alexiadou. 2001. On Stylistic Fronting: Germanic vs Romance LINX. *Working Papers in Scandinavian Syntax* 68, 117-145.

Franco, Irene. 2017. Stylistic fronting in Old Italian: A phase-based analysis. *Language* 93,3, 114-151.

Holmberg, A. 2006. Stylistic fronting. M. Everaert & H. van Riemsdijk (dirs), *The Blackwell Companion to Syntax*. Oxford : Blackwell. 532-565.

Imel, Brock. 2019. *Sa nature proveir se volt*: *A new examination of leftward stylistic displacement in Medieval French through textual domain, information structure, and Oral Représenté*. Thèse de doctorat, University of California Berkeley.

Labelle, Marie et Paul Hirschbühler. 2014. Déplacement stylistique à gauche de verbes non conjugués en ancien et en moyen français. *Corpus* 13, 191-219.

Mathieu, Eric. 2006. Stylistic Fronting in Old French. *Probus* 18, 219-266.

Olivier, M. 2022. Diachronie de la proclise et de l'enclise avec l'infinitif en français médiéval (12e-15e s.). *Studia linguistica*.

Poletto, Cecilia et Francesco Pinzin. En cours. *What's left of the V2 high tide: Infinitival Antepositio*n.

Prévost, Sophie. 2003. Détachement et topicalisation : des niveaux d'analyse différents. *Cahiers de Praxématique* 40, 97-126.

Rahn, Cristina A. 2016. *At the Left Edge: Fronting in Medieval French Embedded Clauses*. Thèse de doctorat, Université de Caen et Universität Konstanz.

Salvesen, C. Meklenborg. 2011. Stylistic Fronting and Remnant movement in Old French. J. Bern, H. Jacobs & T. Scheer (dirs), *Romance Languages and Linguistic Theory. Selected Papers from 'Going Romance' Nice 2009*. Amsterdam: Benjamins. 323-342.

COMMUNICATION C10
L'atlas Dees électronique

Tobias SCHEER & Guylaine BRUN-TRIGAUD
U. Côte d'Azur, Bases, Corpus, Langage (UMR 7320)

La communication présente l'atlas Dees électronique 2022 (ADE22) récemment mis en ligne (http://atlasdees.unice.fr/wordpress). Elle aborde la constitution du corpus qui l'alimente (2,2 Mio de mots, ramenés à 98.230 formes uniques) ainsi que, à l'aide d'un exemple choisi, son fonctionnement et le type de questions auxquelles il peut répondre.

L'atlas Dees 1987 (AD87) est basé sur 200 textes littéraires datés des 13e et 14e siècles et son originalité réside dans le fait qu'A. Dees les a localisés dans l'espace à l'aide de 268 critères linguistiques dégagés par l'étude de chartes (dans son atlas des chartes de 1980). Il a ainsi attribué les textes à des segments d'un maillage géographique de 87 zones, ce qui lui a permis ensuite de produire des cartes montrant la distribution diatopique pour des ensembles de données choisis.

Le propos de l'ADE22 est de rendre vivant et librement accessible l'AD87, i.e. de reconstruire, autant que faire se peut, l'instrument que l'équipe Dees avait mis au point dans les années 80, qui a permis la confection des deux atlas mais n'a jamais été rendu public ni même décrit, et qui est aujourd'hui perdu. Dees et son équipe ont choisi 517 cartes publiées dans l'AD87 selon ce qu'ils jugeaient important. L'AD87 était ainsi figé à tout jamais en fonction des choix et des intérêts de ses concepteurs. Le propos de l'ADE22 est de rendre vivant le travail de Dees et de son équipe, de façon à qu'il soit accessible à tous et permette l'affichage diatopique de données librement choisies. A cette fin, l'utilisateur peut formuler ses propres requêtes dans le corpus des 200 textes, puis afficher la distribution diatopique des résultats.

Le moteur de recherche exploitant le corpus a été fait sur mesure pour le projet. Il permet des recherches par forme et lemme afr., ainsi que par lemme latin. L'input à la requête sont ou bien des chaînes de caractères fixes, ou bien des expressions régulières. La communication explique les choix qui ont été faits lors de la constitution du corpus et de l'alignement progressif du matériau de base fourni par les fichiers informatique Dees d'origine dont nous avons disposé grâce à Piet van Reenen et du NCA : la lemmatisation a été opérée à l'aide du lexique Frolex 3.0, mais demeure largement imparfaite (Frolex ne connaît que 58% des formes uniques). Sur cette base des lemmes reconnus, Frolex permet l'alignement avec les formes latines du FEW (qui ne notent pas la quantité vocalique), qui ensuite ont été à leur tour alignées avec celles du Gaffiot électronique (qui fournit la quantité). Le travail sur l'alignement des formes latines n'est pas achevé : la recherche par forme latine fonctionne mais actuellement présente des défauts majeurs.

Enfin, comme l'AD87, l'ADE22 est un atlas lexical : les parties du discours (avec une grille très fine) sont identifiées par un code à trois chiffres que Dees a associé à chaque mot, mais pour l'instant on ne peut faire des requêtes dans ce code. Par ailleurs, le corpus n'est pas annoté pour d'autres propriétés, notamment morpho-syntaxiques. Toutes ces questions sont autant de pistes pour l'avenir : meilleure lemmatisation, recherche par code Dees, alignement avec le latin, annotation morpho-syntaxique.

Afin de montrer à quel type de questions l'ADE22 peut répondre, une idée de G. Straka est mise à l'épreuve. Lat. -C(i)ca produit une variation concernant le voisement du réflexe de la vélaire : pour °gr<u>a</u>nica par exemple on trouve *grange* autant que *granche*, et cette variation est systématique en afr. L'ayant examinée dans les toponymes, les articles du FEW et une carte de l'ALF (g<u>a</u>llica), Straka (1979 [1970] : 359) conclut que la syncope s'est répandue du Nord-Est vers le Sud (importée par les Francs et suivant leur sillon). Le voisement intervocalique en revanche aurait pris le chemin inverse : venant du Sud, il a gagné du terrain en allant vers le Nord. Ainsi au Sud et à l'Ouest la voisée ʤ a été doublement favorisée : le k d'origine a été atteint par le voisement plus tôt qu'ailleurs, et il a dû

patienter plus longtemps avant d'être figé par la syncope. Il en va de même, à l'inverse, pour ʧ dans le Nord-Est : ici la syncope s'est appliquée de bonne heure, et le voisement s'est fait attendre. La base empirique de cette analyse était limitée par les moyens de l'époque, et l'ADE22 est de nature à confirmer ou réfuter la distribution diatopique conçue par Straka.

CONFÉRENCE C11
La BFM 2022 : un corpus pour les recherches diachroniques en français médiéval et au-delà

Alexey LAVRENTEV
IHRIM (UMR 5317)
Céline GUILLOT BARBANCE
ENS Lyon & IHRIM (UMR 5317)

Depuis plus de 30 ans, la Base de français médiéval (BFM) fournit à la communauté académique un outil précieux pour la recherche sur la langue et la littérature du Moyen Âge. Partie de la numérisation d'éditions de quelques textes symboliques de l'ancien français, elle offre aujourd'hui un corpus de presque 6,5 millions de mots qui s'étend des premiers textes français jusqu'à la fin du XV^{e} siècle.

Seront principalement abordées dans la présentation deux thématiques connexes, la représentativité et l'interopérabilité de la base. La représentativité sur le plan diachronique, ainsi que typologique (genres et domaines textuels) et dialectal, a toujours été un objectif majeur de la BFM. Ses métadonnées sont très riches et précises, ce qui permet de connaître la configuration du corpus et de pallier les éventuels déséquilibres inévitables. TXM, l'outil d'exploitation de la Base permet d'une part de créer des sous-corpus et des partitions sur mesure et, d'autre part, fournit des outils statistiques pour gérer les déséquilibres (spécificités, analyse factorielle des correspondances). Dans la perspective de favoriser son interopérabilité avec d'autres corpus textuels, la BFM s'est aussi très tôt et continuellement engagée dans la standardisation de ses données et métadonnées (encodage XML-TEI, étiquetage morpho-syntaxique *Universal Dependencies*, thésaurus de typologie textuelle du Consortium CAHIER). Cette démarche de standardisation couplée à une politique d'ouverture maximale facilite l'intégration de ses textes dans des corpus diachroniques plus étendus pour le français (GGHF, projet Democrat, intégration avec les Bibliothèques Virtuelles Humanistes) et la création de corpus multilingues comparables (le latin avec le projet Palafra, d'autres langues romanes en perspective).

COMMUNICATION C12
Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval

Sophie PREVOST[1], Loïc GROBOL[2], Mathieu DEHOUCK[1], Alexey LAVRENTEV[3] & Serge HEIDEN[3]
[1] Lattice (UMR 8094, ENS-PSL, U. Sorbonne Nouvelle)
[2] MoDyCo (UMR 7114, U. Paris Nanterre
[3] IRHIM (UMR 5317, ENS de Lyon)

La plupart des changements morpho-syntaxiques et syntaxiques du français se sont déroulés durant la période médiévale. Or, la période du moyen français n'est pas équipée en données enrichies morpho-syntaxiquement et syntaxiquement, et celle de l'ancien français l'est encore insuffisamment. Cette communication a pour objectif la présentation du corpus Profiterole, dont la mise à disposition permet d'envisager un renouveau des études, désormais fondées sur une masse de données significative. Le corpus Profiterole (1 million de mots) a été conçu dans le cadre du projet ANR Profiterole (2017-2022), qui avait trois objectifs : constituer des ressources pour le français médiéval (corpus annoté et lexiques), concevoir des analyseurs syntaxiques pour le français médiéval, et esquisser la modélisation de certains aspects (morpho-)syntaxiques de l'évolution du français.

Cette présentation s'articulera en trois points principaux. Après avoir brièvement rappelé le contexte et les enjeux de la création du corpus Profiterole, nous présenterons les modalités de la constitution de la ressource à enrichir, tant en ce qui concerne le choix raisonné des textes (10^{e} -13^{e} s.), tous issus de la Base de Français Médiéval, que le choix d'échantillonner les textes afin d'accroître la diversité des données. Dans un second temps nous présenterons les modalités d'annotation du corpus en parties du discours et en dépendances syntaxiques dans le cadre du projet Universal Dependencies (UD). Nous exposerons les deux démarches parallèles qui ont été menées : l'utilisation d'un analyseur syntaxique symbolique et celle d'analyseurs à base de réseaux neuronaux. Ces derniers se sont appuyés sur la ressource existante SRCMF (Syntactic Reference Corpus of Medieval French, http://srcmf.org/), corpus d'ancien français de 250 000 mots étiquetés en POS et annotés en dépendances syntaxiques, converti en 2018 au format UD. À l'issue de l'annotation par les différents analyseurs, un algorithme de vote a été mis en œuvre pour prédire les étiquettes syntaxiques et les attachements les plus probables. S'ensuit une correction manuelle d'une partie des données annotées, point de départ d'un processus de bootstrapping : ré-entraînement des analyseurs sur les données corrigées (« gold »), analyses des données non vérifiées, application de l'algorithme de vote, corrections d'une partie des données, ré-entrainement, etc. L'objectif final est d'obtenir un corpus entièrement vérifié. Dans un troisième temps nous évoquerons la distribution du corpus (projet UD, Ortolang et HumaNum), et ses modalités d'exploration/interrogation via la plateforme TXM (http://txm.ish-lyon.cnrs.fr/bfm/), qui offre désormais la possibilité de requêtes syntaxiques.

Références

Grobol, L., Prévost, S. et Crabbé, B. (2021). Is Old French tougher to parse? In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories* (TLT, SyntaxFest 2021), Sofia, Bulgaria. Association for Computational Linguistics, 27-34.

COMMUNICATION C13
Un exemple de corpus annoté en diachronie longue : le corpus Democrat, enjeux et exploitations

Julie GLIKMAN[1], Frédéric LANDRAGIN[2], Catherine SCHNEDECKER[1] & Amalia TODIRASCU[1]
[1] U. de Strasbourg, LiLPa (UR 1339)
[2] Lattice (UMR 8094)

Le corpus DEMOCRAT, constitué dans le cadre du projet ANR de même nom[1], dont il représente l'un des livrables, est un corpus de textes annotés en chaînes de référence. Chaque élément référentiel (un 'maillon') est indiqué par un identifiant rapportant à son référent, ce qui permet la création d'une chaîne rassemblant toutes les mentions d'un même référent. En outre, le corpus bénéficie également d'une annotation en partie du discours. Ce corpus a déjà fait l'objet de plusieurs exploitations (voir notamment Schnedecker *et al.* (dir.) 2017 ; Landragin (dir.) 2021). À l'occasion de ce colloque, nous souhaitons mettre en avant les défis méthodologiques en jeu lors de la constitution d'un tel corpus. Nous montrerons ainsi que, malgré les difficultés inhérentes à un tel projet, les exploitations du corpus Democrat en diachronie longue permettent d'obtenir des résultats convaincants.

La constitution d'un corpus en diachronie longue doit, pour garder son intérêt, permettre son exploitation diachronique, à travers des éléments comparables. Cela suppose, d'une part, la nécessité de s'interroger dès le choix des textes du corpus, en termes de représentativité tant au niveau des genres qu'au niveau des périodes (cf. Landragin 2021[2] et Prévost 2020). D'autre part, cela suppose également d'opter pour des choix d'annotation pouvant s'appliquer sur les différents états de langue[3]. Le corpus DEMOCRAT permet ainsi d'interroger la nature et la constitution des chaînes de référence en diachronie : nature des maillons, nombre de maillons, longueur des chaînes, densité référentielle, distance intermaillonaire, nombre de référents dans un texte, nature des premières mentions, etc. Outre des données relevant de l'évolution connue du français (diminution des sujets nuls par ex.), l'exploitation du corpus nous a permis d'obtenir des résultats sur des comparaisons intra-génériques (sur les récits narratifs brefs, cf. Obry *et al.* 2017, ou sur les textes encyclopédiques, cf. Oberlé *et al.* 2018), mais aussi sur l'évolution de la structuration textuelle (cf. Capin *et al.* 2021). L'exploitation du corpus Democrat nous amène ainsi à une meilleure compréhension non seulement du fonctionnement des chaînes de référence, mais aussi des genres textuels, et de leur évolution diachronique respective.

[1] ANR-15-CE38-0008, 2016-2020, rassemblant des équipes notamment des laboratoires Lattice (Paris), LiLPa (Strasbourg), ICAR et IHRIM (Lyon). Site du projet : https://www.lattice.cnrs.fr/democrat/. Le corpus est distribué sur Ortolang : Frédéric Landragin. Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, AFIA, 2016, pp.11-15. ⟨afia.asso.fr⟩. ⟨hal-01347949⟩
Langues, textes, traitements informatiques, cognition - UMR 8094 (Lattice), Linguistique, Langues, Parole - EA 1339 (LiLPa), Interactions, corpus, apprentissages et représentations - UMR 5191 (ICAR), Institut d'histoire des représentations et des idées dans les modernités - UMR 5317 (IHRIM) (2019). Democrat [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v1.1, https://hdl.handle.net/11403/democrat/v1.1.
[2] Voir la liste des textes du corpus : https://www.lattice.cnrs.fr/democrat/files/ANR-15-CE38-0008-DEMOCRAT_livrable_L1_corpus.pdf.
[3] Voir Livrable 2 : Manuel d'annotation du corpus, https://www.lattice.cnrs.fr/democrat/files/ANR-15-CE38-0008-DEMOCRAT_livrable_methodo.pdf.

Références

Capin, D., Glikman, J., Schnedecker, C. & Todirascu, A. (2021). Le rôle des chaînes de référence dans la structuration textuelle : étude diachronique de l'ancien français au français moderne. *Langages*, 224, 87-107. https://doi.org/10.3917/lang.224.0087

Landragin F. (dir.) (2021) Un corpus annoté en chaînes de référence et son exploitation : le projet DEMOCRAT. *Langages* 2021/4 n°224.

Landragin, F. (2021). Le corpus DEMOCRAT et son exploitation. Présentation. *Langages*, 224, 11-24. https://doi.org/10.3917/lang.224.0011

Oberlé, B., Schnedecker, C., Baumer, E., Capin, D., Glikman, J., Guo, C. & Tushkova, J. (2018). Les chaînes de référence dans les textes encyclopédiques du 12e au 21e siècle : étude longitudinale. *Travaux de linguistique*, 77, 67-141. https://doi.org/10.3917/tl.077.0067

Obry, V., Glikman, J., Guillot-Barbance, C. & Pincemin, B. (2017). Les chaînes de référence dans les récits brefs en français : étude diachronique (XIIIe-XVIe s.). *Langue française*, 195, 91-110. https://doi.org/10.3917/lf.195.0091

Prévost S. (2020) Une grammaire fondée sur un corpus numérique. In Marchello-Nizia C., Combettes B., Scheer T. & Prévost S (2020). *Grande Grammaire Historique du Français* (GGHF). De Gruyter, p. 37-53.

Schnedecker C., Glikman J. & Landragin F. (dir.) (2017) *Les chaînes de référence en corpus*. *Langue Française* 2017/3 n°195.

COMMUNICATION C14
De l'utilité de confronter les sorties de plusieurs étiqueteurs morphosyntaxiques

Adam RENWICK
UGA, LiDiLEM

De nombreux logiciels ont été développés pour faciliter l'étiquetage morphosyntaxique de corpus, fonctionnant à l'aide des méthodologies variant des approches stochastiques en passant par des modèles Markov cachés aux approches neuronales. Lorsque ces méthodologies sont combinées avec les avancées en informatique, il est possible d'étiqueter des milliers de mots par seconde sur un quelconque ordinateur, mais le chercheur est néanmoins confronté non seulement au choix d'un outil parmi plusieurs, mais aussi à l'identification du juste équilibre entre ses capacités (en programmation, par exemple), ses contraintes (de temps, de budget, de capacité de calcul informatique) et la fiabilité de l'étiquetage (morphosyntaxique, des dépendances...).

Certains outils, tels que Deucalion Old French (Camps, Clérice et Pinche 2019), l'interface UDPipe2 de Straka (2018), SEM (Tellier, Dupont et Courmet 2012) ou CoreNLP (Stanford NLP Group 2022a), permettent un étiquetage via une simple interface en ligne, ne nécessitant que quelques clics pour étiqueter un corpus. Alors que la simplicité des pages web peut comporter des contraintes si de nombreux textes sont à traiter ou si des limites existent quant à leur longueur, la prolifération de modèles pré-entraînés sur des langues/états de langue – Stanza fournit, par exemple, plus de 90 modèles pré-entrainés sur 66 langues (Stanford NLP Group 2022b) – permet de contourner ces difficultés sans que des connaissances informatiques ne deviennent trop importantes. Diamétralement opposée à cette première approche de simplicité, se dresse la possibilité de créer de toutes pièces chacun des éléments nécessaires à la création d'outils pour étiqueter les propriétés morphosyntaxiques des mots d'un corpus, les éléments nécessaires comprenant par exemple BERT (Devlin et al 2018), CamemBERT (Martin et al 2020) et FlauBERT (Le *et al.* 2020), des corpus sous forme et en quantité adéquates pour l'entraînement des modèles aussi bien que pour la création des modèles eux-mêmes), avant d'aborder les capacités computationnelles, budgétaires et en programmation pour mener à bien la création d'un outil performant. Ces limites ont pour résultat qu'un tel flux de travail n'est pas à la portée de tout chercheur ou équipe de recherche.

Dans le cadre de cette communication, nous aborderons le cas d'un corpus de français s'étalant sur quatre siècles dont les états de langue impliquent d'importants changements sur les plans de la syntaxe et du lexique. Nous analyserons les possibilités qu'offre une analyse qui cherche un juste équilibre entre ces extrémités, en analysant la faisabilité et la fiabilité de l'emploi simultané de plusieurs étiqueteurs dans le but de confronter leurs résultats. Nous démontrerons que la barrière informatique à l'harmonisation de ces analyses est bien moins importante que celle impliquée dans la création et l'entraînement de modèles et ce, tout en restant moins chronophage et énergivore. En abordant la précision des résultats issus des confrontations d'étiquettes morphosyntaxiques, nous contextualiserons également la précision et l'utilité des différents étiquetages proposés lorsque les textes à analyser ne correspondent pas précisément à ceux utilisés dans la création des outils selon le processus *test-train-validate*.

Références

CAMPS Jean-Baptiste, CLÉRICE Thibault et PINCHE Ariane (2019). *Deucalion, Modèle Ancien Francais (0.2.0)*. École Nationale des Chartes. https://doi.org/10.5281/zenodo.3237455

DEVLIN Jacob, CHANG Ming-Wei, LEE Kenton et TOUTANOVA Kristina (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805v2

LE Hang, VIAL Loïc, FREJ Jibril, SEGONNE Vincent, COAVOUX Maximin, LECOUTEUX Benjamin, ALLAUZEN Alexandre, CRABBÉ Benoît, BESACIER Laurent, SCHWAB Didier. (2020). FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français Dan.s *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).* Volume 2 : Traitement Automatique des Langues Naturelles. p. 268-278.

MARTIN Louis, MULLER Benjamin, ORTIZ SUÁREZ Pedro Javier, DUPONT Yoann, ROMARY Laurent, VILLEMONTE DE LA CLERGERIE Éric, SEDDAH Djamé et SAGOT Benoît (2020). CamemBERT : a Tasty French Language Model. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p.7203–7219.

Stanford NLP Group (2022a). Core NLP Demo. https://corenlp.run

Stanford NLP Group (2022a). Available Models & Languages. https://stanfordnlp.github.io/stanza/available_models.html

STRAKA, Milan (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. Dans *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, p. 197–207. https://ufal.mff.cuni.cz/udpipe/2

TELLIER Isabelle, DUPONT Yoann et COURMET Arnaud. (2012). Un segmenteur-étiqueteur et un chunker pour le français. Dans *Actes de la conférence conjointe JEP-TALN-RECITAL 2012.* Volume 5 : Démonstrations. p. 7-8. http://apps.lattice.cnrs.fr/sem/

SESSION 4 – Effectuer des recherches avec les corpus constitués

COMMUNICATION C15
Observations diachroniques dans un corpus de presse avec le Lexicoscope

Sascha DIWERSY
U. Paul Valéry Montpellier 3, PRAXILING (UMR 5267)
Olivier KRAIF
UGA, LiDiLEM

Dans cette communication, nous présentons les nouvelles fonctionnalités diachroniques d'une plate-forme d'interrogation de corpus analysés en dépendances nommée Lexicoscope (Kraif & Diwersy ; 2012 ; Kraif 2016 ; Kraif 2019). Cette plate-forme vise à fournir, en plus des outils standards fournis en linguistique de corpus (concordances, co-occurrences, listes de fréquences, extraction de mots-clés, recherche de patterns), des fonctionnalités avancées d'interrogation syntaxique, tout en restant à la portée d'un utilisateur non initié à la manipulation de langages de requêtes complexes, en s'appuyant sur le requêtage basé sur l'exemple (Agostinus *et al.*, 2016).

Différents corpus sont actuellement installés sur le Lexicoscope, dont certains comportent une dimension diachronique : c';est le cas du corpus Phraseorom, un grand corpus de romans contemporains en 3 langues (français, anglais, allemand) s'étalant de 1950 aux années 2010 dédié à l'étude contrastive des sous-genres littéraires (Diwersy *et al.*, 2021), d'un corpus de romans français du 19e siècle couvrant la période 1830-1899, équilibré par décennies (Sorba 2022), ainsi que d'un corpus de presse nommé *lmdiachro*, rassemblant des articles de 1944 à nos jours. Un autre corpus romanesque est en cours de préparation, constitué de romans de chevalerie à travers divers états de langue, du moyen-âge au 17e siècle (Coavoux *et al.*, 2022).

Pour ces corpus constitués dans une perspective diachronique, de nouvelles fonctionnalités ont été ajoutées récemment, afin d'afficher le profil chronologique d'une expression prise globalement ou considérée à travers ses réalisations variables et ses diverses co-occurrences. Ces fonctionnalités vont du chronogramme aux indicateurs de tendances, en passant par la clusterisation automatique de période (Gries & Hilpert, 2008).

L'objectif de cette communication est de présenter ces fonctionnalités à travers le corpus de presse *lmdiachro*, récemment construit, qui propose un échantillon d'articles du quotidien *Le Monde*, années après années, de 1944 à 2015. Pour des raisons de droits d'auteur, le corpus ne peut pas être diffusé tel quel, mais peut être interrogé grâce aux outils de concordance et d'extraction statistique du Lexicoscope.

Après avoir brièvement présenté les fonctionnalités du Lexicoscope, nous détaillerons la constitution du corpus *lmdiachro*, et aborderons les différentes méthodes dédiées à l'analyse chronologique de corpus. Nous illustrerons ensuite ces nouvelles fonctionnalités d'exploration chronologique à travers une étude de cas basée sur des observations tirées du corpus *lmdiachro*. Enfin, nous exposerons les perspectives de développement futur pour l'analyse diachronique.

Références

Augustinus, L., Vandeghinste, V., Vanallemeersch, T. (2016). Poly-GrETEL: Cross-Lingual Example-based Querying of Syntactic Constructions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3549–3554, Portorož, Slovenia. ELRA.

Coavoux M., Denoyelle, C. Kraif, O., Sorba J. (2022) Phraséologie du roman médiéval en prose. *Colloque international DIACHRO X, Le français en diachronie*, Paris, Sorbonne Université, juin 2022.

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

Gries S. Th. and Hilpert M. (2008). The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, 3, 59–81.

Kraif O. (2016). Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. *Cahiers de lexicologie* 108 : 91-106.

Kraif O. (2019). Explorer la combinatoire lexico-syntaxique des mots et expressions avec le lexicoscope, *Langue française* 203 : 67-82. DOI : 10.3917/lf.203.0067

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

Sorba J. (2022). *Phraséologie et genres textuels. Perspectives synchroniques et diachroniques*. Habilitation à diriger des recherches, mémoire de synthèse. Université Grenoble Alpes.

COMMUNICATION C16
A diachronic corpus to study modality in the Latin language: the WoPoss experience step by step

Francesca Dell'Oro & Helena Bermúdez Sabel
U. Neuchâtel, Suisse

The main aim of the WoPoss project is to reconstruct the evolution of a selection of modal markers (see the list in the Guidelines [Dell'Oro 2022]) through the history of the Latin language, covering one millennium from the 3rd century BCE to the 7th century CE. While diachrony is the main variable, we also aim at representativity in terms of other linguistic criteria, taking into account both documentary and literary attestations in order to set up a balanced corpus (Cuzzolin - Harverling 2009). In fact, one of the goals of the project is to investigate by which corpus variables and which kinds of linguistic variation in the Latin diasystem (Coseriu 1969) the use of modality is influenced or even determined. In this paper we present each of the main stages which brought us to the setting up of the WoPoss corpus, including the design of the annotation schemes and the development of the search interface (https://woposs.unine.ch/form.html). We also show how the corpus can be used to study modality (defined according to Nuyts 2016).

We outline step by step

- the conception of the annotated corpus from its design to its publication as a TEI-compliant corpus that contains (1) the lemma and part-of-speech of each token, (2) the linguistic description of each modal passage including the identification of all its constituents, (3) structural markup in conjunction to editorial interpretations, (4) metadata about the sources, the work and its author;
- the selection of texts and the criteria applied (chronology, genre, open source availability);
- the elaboration of the guidelines for automatic and manual annotation and the implementation of the work of the annotators in the Inception platform (Klie *et al.* 2018);
- the conception of the search interface and its development (see Figures 1 and 2);
- the exploitation of the corpus: we show which markers are more stable diachronically, which syntactic changes took place and when and how modality-related phenomena can be connected to genre;
- and finally, the improvements we plan for the future.

We address the challenges we overcame for each step and we suggest some solutions in order to be able to guarantee the correct analysis of the results when there was no optimal solution. We hope that the outlined use case could be useful also to other projects working on other languages and/or other diachronic phenomena.

Références

Coseriu, E. 1969. *Einführung in die strukturelle Linguistik*. Tübingen: G. Narr/R. Windisch.

Cuzzolin, P. & Haverling, G. 2009. Syntax, sociolinguistics, and literary genres, In: *NewPerspectives on Historical Latin Syntax*, Volume 1, Edited by Ph. Baldi & P. Cuzzolin, Berlin -New York: Mouton de Gruyter, pp. 19-64.

Dell'Oro, F. 2022. *WoPoss guidelines for annotation. Revised version*. Zenodo. https://doi.org/10.5281/zenodo.6417878

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: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, Santa Fe: Association for Computational Linguistics, pp. 5-9. http://tubiblio.ulb.tu-darmstadt.de/106270/

Nuyts, J. 2016. Analyses of the modal meanings. In: *The Oxford Handbook of Modality and Mood*, Edited by J. Nuyts & J. van der Auwera. Oxford: Oxford University Press, pp. 31-49. https://doi.org/10.1093/oxfordhb/9780199591435.013.1

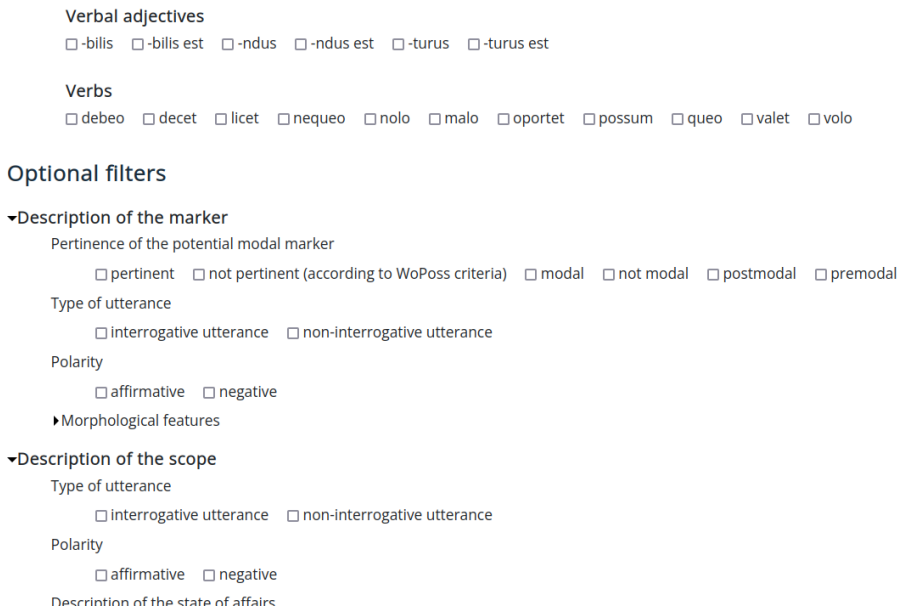


Fig. 1. Screenshot of a section of the faceted search form

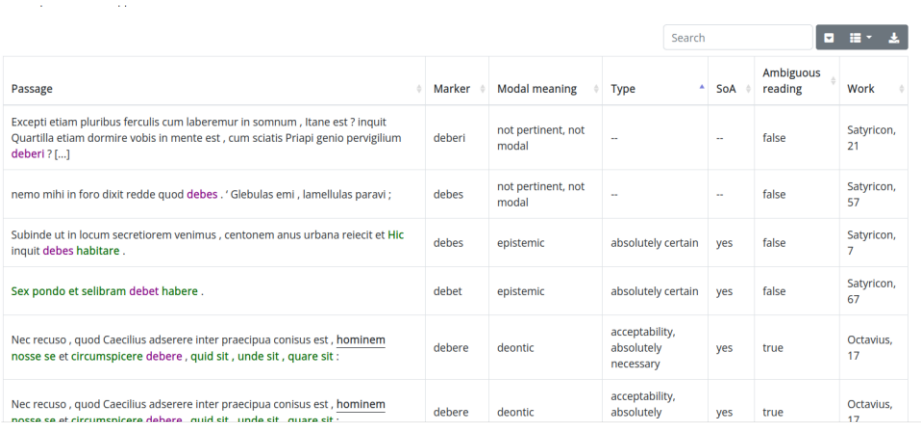


Fig. 2: Screenshot with some of the results obtained after performing a query by marker (*debeo*)

Acknowledgements

This work was supported by the Swiss National Science Foundation (SNSF N° PP00P1_176778) and it is led by Francesca Dell'Oro at the University of Neuchâtel.

COMMUNICATION C17

La variation terminologique en musique : l'harmonie à la loupe de la textométrie

Eleonora MARZI
Università di Bologna, Département de Langue, Littérature et Cultures modernes, Italie

Parmi les nombreuses perspectives de recherche qui s'ouvrent grâce à la disponibilité de corpora en format numérique en accès ouvert (Clarin[1], Ortolang[2] parmi d'autres) ou grâce à la possibilité d'en construire de manière relativement rapide à travers des plateformes en accès libre (Gallica.fr.[3], Open Library[4] parmi d'autres), celle de la terminologie diachronique montre tout son intérêt si l'on considère l'exploitation de grandes masses de données diversifiés en termes de langue de spécialité couvrant des larges intervalles temporels. Persuadés que l'étude de l'évolution lexicale des concepts reflète et parfois même suggère les changements que la société traverse, notre contribution porte sur la modalité de construction et d'analyse d'un corpus de langue de spécialité en musique qui répond à des besoins de recherche en terminologie notamment sur le repérage de la variation sémantique dans une perspective diachronique.

Pour exemplifier le parcours de conception, construction et analyse nous présenterons le cas d'étude du mot *harmonie* qui du XVII[e] siècle au XXI[e] siècle modifie légèrement son sens. Bien que dans la langue généraliste le mot *harmonie* indique généralement une combinaison de plusieurs parties dans un ensemble : « [En parlant de ce qui est perçu par l'oreille ou par l'œil] Combinaison spécifique formant un ensemble dont les éléments divers et séparés se trouvent reliés dans un rapport de convenance, lequel apporte à la fois satisfaction et agrément » (TLFi – Trésor de la Langue française informatisé) », dans la langue de spécialité musicale le sens change suivant l'histoire de la musique.

Après avoir présenté la méthodologie pour la conception et la collecte d'un corpus monolingue spécialisé sur la musique couvrant un intervalle de temps du XVII[e] siècle au XXI[e] nous l'analyserons grâce aux mesures de la textométrie qui implique l'application de l'analyse statistique aux textes – donc quantitative, tout en permettant un retour aux cas spécifiques – donc qualitative grâce au logiciel en accès libre TXM[5].

Notre contribution a donc un double objectif : 1) expliciter la méthodologie de construction d'un corpus musical qui soit représentatif d'une hypothèse de recherche sur la variation sémantique en terminologie diachronique et 2) repérer les variations sémantiques du mot *harmonie* à travers l'analyse de son comportement syntaxique et sémantique par le biais des calculs textométriques.

[1] https://www.clarin.eu/
[2] https://www.ortolang.fr/en/home/
[3] https://gallica.bnf.fr/
[4] https://openlibrary.org/
[5] Heiden S., Magué J.-P., Pincemin B., 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* (pp. 12 p.). Rome, 2010. Retrieved from http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf

Références

Altmanova J. Zollo S.D. (éd.) La néologie à l'ère de l'informatique et de la révolution numérique. *Neologica*, 13, Classiques Garnier, Paris, 2019

Balnat V. Gérard C. (éd.), Les études de néologie au XXI[e] siècle. Un état de la recherche européenne. *Neologica*, 15, Classiques Garnier, Paris, 2021

Biber D. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4) : 243-257, 1993

Deliège C. Approche d'une sémantique de la musique. *Revue belge de Musicologie / Belgisch Tijdschrift voor Muziekwetenschap*, 1(4) : 21-42, 1966

Heiden S., Magué J.-P., Pincemin B., TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In JA*DT 2010 : 10th International Conference on the Statistical Analysis of Textual Data* (pp. 12 p.). Rome, 2010. Retrieved from http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf

Humbley J., *La néologie terminologique*, Éditions Lambert-Lucas, Limoges, 2018

Humbley, J. (éd.), *Multiple perspectives on terminological variation*, John Benjamins, Amsterdam/Philadelphia, 2017, pp.181-212

Mayaffre D., Pincemin B., Poudat C. Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse des discours. *Langue française, Les outils informatiques au service des linguiste*s, Armand Colin, 2019, pp.101-115

Pernon G. *Dictionnaire de la musique*, Gisserot, Brest, 2007

Pincemin B, Guillot C., Heiden S. *et al.*, Usages linguistiques de la textométrie. Analyse qualitative de la consultation de la Base de Français Médiéval via le logiciel Weblex, *Syntaxe et Sémantique*, 9 (1) : 87-110, 2008 DOI : 10.3917/ss.009.0087

Pincemin B. Sémantique interprétative et textométrie, *Corpus*, 10 : 259-269, 2011 DOI : 10.4000/corpus.2121

Labbé D, Le calcul du sens des mots. La lexicologie assistée par ordinateur. *Mathématiques et société*, Université de Neuchâtel, Neuchâtel, 2010.