Introduction

A l'origine, ce recueil était destiné à être publié dans le cadre d'une section du congrès de l'Association allemande de francoromanistes (FRV) qui aurait dû se tenir à l'Université de Vienne en février 2020. L'objectif de cette section, intitulée « Approches numériques des corpus historiques de langues de France », était de prendre en compte et de discuter les différentes approches méthodologiques de la linguistique de corpus pour l'exploitation et l'analyse des niveaux linguistiques anciens du français et de ses variétés diasystématiques, ainsi que des autres langues de France (par exemple l'occitan, le catalan, etc.). Le congrès a finalement été annulé en raison de la pandémie de coronavirus qui s'annonçait, c'est pourquoi nous avons voulu rassembler les résultats scientifiques de la section dans un recueil. En outre, nous avons décidé de lancer un nouvel appel à communications afin d'atteindre d'autres chercheurs travaillant dans le domaine de la linguistique de corpus historique et utilisant des instruments numériques. Au total, le volume réunit huit contributions de chercheurs d'Allemagne, de France et du Canada, qui présentent les résultats de leurs recherches.

L'accent est notamment mis sur les méthodes appliquées. La linguistique de corpus, avec ses paradigmes de recherche, s'y prête particulièrement bien : Depuis le début de la recherche en linguistique de corpus dans la seconde moitié du 20ème siècle, il s'agissait surtout de recherches synchrones (principalement sur l'anglais). Pour la nouvelle discipline qui se développe et qui vise à la reconnaissance automatique de textes, des exemples de textes actuels et proches des standards constituent la base de données naturelle. Les progrès de la technologie de l'information et la numérisation croissante qui en découle ouvrent également de nouvelles voies de recherche méthodologique et empirique aux sciences humaines et à la linguistique en particulier. Alors que les procédés numériques se sont établis depuis longtemps dans la recherche linguistique moderne, entre autres dans la phonétique expérimentale, acoustique ou perceptive (par ex. avec *praat* ou *SpeechRecoder*), les corpus historiques en tant que partie de la recherche linguistique représentent encore à bien des égards un défi pour les sciences humaines numériques. Il ne faut pas sous-estimer la tâche que représente par exemple la numérisation de manuscrits anciens, malgré la maîtrise des logiciels OCR et leur facilité d'utilisation,

par exemple *OCR4all* à l'Université de Würzburg en Allemagne (Reul et al. 2019 ; Wehner et al. 2020).[1]

Outre les grands projets de corpus tels que le corpus *Frantext*, dont la base de données est essentiellement constituée de textes récents, la création[2] et l'exploitation de corpus de textes en ancien français à l'aide de méthodes numériques ont commencé très tôt dans le domaine francoromanistique, comme, par exemple, dans le projet commun *Sources et outils pour l'analyse du français ancien* (SOFA ; cf. Stein/Glessgen 2005) aux universités de Stuttgart, Zurich et Ottawa. Parallèlement, Guillot et al. soulignent qu'il existe certes une quantité considérable de corpus pour l'ancien et le moyen français,[3] mais qu'ils ont souvent été consacrés à un usage très spécifique et n'étaient pas non plus librement accessibles :

> Une spécificité des corpus de français du Moyen Age est leur nombre, relativement élevé, et leur dispersion tant dans l'espace que dans les objets qu'ils se sont fixés. Il est des raisons à ce polycentrisme : jusqu'en 2005 environ, plusieurs des corpus étaient en quelque sorte « privés », et donc difficilement accessibles aux chercheurs extérieurs ou étrangers au laboratoire qui les avait créés. D'autre part, le droit protégeant les intérêts des éditeurs commerciaux français étant particulièrement strict, l'ouverture de ces corpus n'était guère encouragée. Enfin, plusieurs corpus se sont centrés sur une période (Base textuelle du moyen français) ou sur des textes reflétant la langue d'une certaine aire géographique (Anglo-norman Hub) ou sur un auteur et/ou la tradition manuscrite d'un texte (projet Charrette, projet Christine de Pizan). Le chercheur qui voudrait utiliser en 2008 un corpus d'ancien ou de moyen français a donc à sa disposition, selon son orientation ou ses besoins, plusieurs corpus bien différenciés. L'envers de cette richesse et de cette possibilité de choix est qu'il n'existe pas actuellement de corpus de référence pour ces périodes (2008, 1).

À côté de la disponibilité numérique, se pose alors, dans un deuxième temps, la question de la fonction, fondamentale en linguistique de corpus, de l'interrogation automatisée des formes linguistiques. En France, un logiciel d'analyse textuelle – *TXM* (Heiden/Magué/Pincemin 2010) – a été développé, notamment à l'Université de Lyon, qui peut également être utilisé pour des corpus historiques, comme par exemple la *Base de Français Médiéval* (Guillot-Barbance/Heiden/Lavrentiev 2017), qui

---

[1] Cf. aussi pour la numérisation d'imprimés français du XVII[e] siècle par OCR Gabay/Clérice/Reul (2023).

[2] Cf. à ce sujet la genèse de la *Base de Français Médiévale* (BMF) à la fin des années 1980 : http://bfm.ens-lyon.fr/spip.php?article339 (dernier accès le 24.9.2024).

[3] Guillot et al. fournissent un aperçu des corpus existants dans leur article (2008, 2-5).

comprend 170 textes du IXe au XVe siècle (http://txm.bfm-corpus.org/). En outre, des projets à orientation historique sur la variation diasystématique du français – avec des originaux numérisés (cf. CHSF : *Corpus Historique du Substandard Français*, Thun 2011 & 2018) – témoignent de la pertinence d'une telle perspective de recherche. Enfin, l'importance des études numériques peut être mesurée par le fait qu'elles marquent, selon Guillot et al., l'entrée dans une nouvelle ère de recherche sur les états de langue plus anciens :

> En linguistique en particulier, on ne peut plus faire l'économie d'une approche « corpus-sielle » ; de ce point de vue nous sommes entrés, épistémologiquement, dans un nouvel âge. En effet, ce nouveau type d'approche permet – contraint à – deux avancées capitales dans le champ des analyses linguistiques.
>
> D'une part les « observables » linguistiques sont objectivés par 1) une explication au sein du corpus de ce qui est prise en compte dans l'analyse […], 2) le paramétrage, au moyen de descripteurs explicites, du périmètre sociolinguistique de ce qui est pris en compte dans l'analyse […], 3) l'échange des observables grâce à la compatibilité désormaisrendue [sic] possible, et le cumul des acquis par le biais de l'échange ou de l'accès aux mêmes corpus. […] D'autre part – et c'est là un point essentiel dans les études diachroniques – la quantification des faits langagiers devient opératoire : désormais, la récurrence d'apparition, ou d'absence, peut prendre une valeur heuristique voire probatoire pour certaines études linguistiques (2008, 8).

Les articles réunis dans ce recueil donnent une très bonne vue d'ensemble pour évaluer dans quelle mesure le potentiel des méthodes de recherche numériques pour le français et les langues de France peut être exploité dans le domaine de la linguistique historique. L'accent est mis sur différents aspects, tels que

- différents projets de numérisation de corpus de langues anciennes, du Moyen Âge à l'époque moderne, et les défis qui en découlent lors de la création et de la préparation des corpus, ainsi que les solutions correspondantes (Goux/Decoux/Pica ; Röttgermann ; Couffignal ; Reichle et al.)
- la description des défis liés au travail avec des corpus existants, par exemple *Frantext* ou la *BFM*, sur des thèmes linguistiques (Tmart ; Schöntag ; Hesselbach)
- la présentation d'outils numériques pour le traitement numérique de textes de niveaux linguistiques plus anciens (Premat/Poggio)

Dans leur contribution « Restituer des fragments de phonologie avec le PAM (*Programme d'Analyse Métrique*) : l'élision de schwa et des voyelles

pleines dans les monosyllabes fonctionnels en très ancien français », **Timothée Premat** (Paris) et **Enzo Poggio** (Montréal) montrent, à l'aide du programme *PAM* (*Programme d'Analyse Métrique*), dans quelle mesure la linguistique informatique et la phonologie diachronique peuvent profiter l'une de l'autre de manière synergique. Ce programme doit faciliter l'étude phonologique de textes en moyen français en les scannant et en les taguant selon leur type prosodique. Concrètement, il s'agit de savoir comment se produit l'élision non transcrite au niveau syllabique pour les voyelles « e », « i » et « o ». A l'aide d'une analyse statistique de la *Chanson de Roland* et de la *Vie de saint Alexis*, il est montré que l'élision non transcrite ne concerne pas seulement les monosyllabes fonctionnels en schwa. D'autres voyelles sont concernées, même si elles s'avèrent plus résistantes à l'élision.

Mathieu Goux (Caen), **Prune Decoux** (Artois-Bordeaux) et **Morgane Pica** (Lyon) présentent dans leur article « Le corpus ConDÉ : bilans, leçons, perspectives » le corpus ConDÉ : un corpus juridique qui documente l'histoire du droit normand du Moyen Âge à l'époque moderne. Le corpus comprend environ 30 millions de signes et se compose de dix textes qui ont été adaptés sur le plan structurel, linguistique et thématique. Dans leur contribution, les auteurs réfléchissent au processus de création du corpus et abordent les phases de sélection des textes et d'encodage. Enfin, ils abordent également les difficultés rencontrées et les solutions possibles.

Julia Röttgermann (Trèves) présente dans son article « Établissement d'un corpus de romans français du XVIII^e^ siècle dans le cadre du projet *Mining and Modeling Text* » son corpus créé dans le cadre du projet interdisciplinaire *Mining and Modeling Text*, qui se compose de 200 romans français des années 1750-1800. En comparaison avec les autres contributions, le corpus comprend ainsi exclusivement des textes littéraires. Ceux-ci sont entre autres enrichis de métadonnées dans les domaines de la narrativité et du genre. Conformément aux principes des données de recherche FAIR, toutes les données sont mises à disposition gratuitement en ligne et peuvent ainsi être utilisées pour d'autres études scientifiques.

Zeïna Tmart (Lyon) travaille principalement sur le fondement de la *Base de Français Médiéval* et veut établir un corpus annoté et représentatif des états de langue du XII^e^ au XVI^e^ siècle. L'intérêt linguistique de la contribution est de décrire l'influence de la grammaticalisation des articles définis et indéfinis sur la structure syntaxique des syntagmes nominaux coordonnés en français.

Gilles Couffignal (Paris) s'est penché sur la langue régionale de l'occitan dans son article « Philologie numérique et données bruitées : un exemple de recherche sur l'occitan prémoderne ». Il se concentre sur l'analyse numérique de textes imprimés en occitan du XVI^e^ siècle. Il combine les méthodes philologiques traditionnelles et les possibilités du travail sur corpus numérique. La contribution est exemplaire pour les variétés qui ont été peu étudiées jusqu'à présent et qui fournissent une base de données relativement petite.

Roger Schöntag (Mannheim) utilise des corpus déjà existants comme base empirique pour vérifier sa question de recherche : dans son article « Le renforcement pittoresque de la négation en ancien français – Possibilités et limites d'une enquête en ligne », il étudie la négation en ancien français et aborde ses origines en latin vulgaire. Le corpus *Frantext* sert de base à l'élaboration de résultats empiriques. Dans sa contribution, l'auteur aborde les difficultés rencontrées lors de l'étude du corpus, dues par exemple à la complexité des formes de négation en ancien français comme la « négation pittoresque », et à la variance orthographique médiévale.

Dans son article « Perspectives de linguistique de corpus sur les formations superlatives latines (-ISSIMUS) dans l'histoire du français », **Robert Hesselbach** (Erlangen-Nürnberg) se concentre sur la question de savoir si et dans quelle mesure le français échappe, dans le domaine du superlatif, à une évolution que l'on peut observer dans d'autres langues romanes, comme par ex. en portugais, en espagnol et en italien. L'auteur examine trois périodes différentes de l'histoire de la langue française et peut démontrer, à l'aide d'une analyse de corpus (de *Frantext*), que ce phénomène n'est attesté en français que pour un certain nombre d'adjectifs. Il est intéressant de noter que les données du corpus permettent de conclure que l'emploi de *-issime* est en augmentation depuis l'époque de l'ancien et du moyen français et qu'il est très populaire en français contemporain dans certains genres de textes ou traditions discursives pour la formation de néologismes.

Elisabeth Reichle (Ratisbonne), **Maria Selig** (Ratisbonne), **Sebastian Ortner** (Munich) et **Rembert Eufe** (Tubingen) présentent dans leur article « PaLaFra : Un projet numérique pour mieux comprendre le passage du latin à l'ancien français » le subcorpus *PaLaFraLat*, qui doit représenter la période de transition du latin tardif au début de l'ancien français et qui a été créé dans le cadre du projet *PaLaFra* entre 2015 et 2018. Concrètement, *PaLaFraLat* consiste en un corpus de textes méro-

vingiens lemmatisés et annotés morphosyntaxiquement, qui doit, entre autres, montrer une approche détaillée et systématique de la variation graphique et morphologique du latin et permettre la comparaison de plusieurs subcorpus pouvant être organisés différemment. L'article illustre cette démarche par une étude exemplaire de textes mérovingiens.

*

Cette publication n'aurait pas été possible sans le soutien de différents collègues et institutions. Nous voulons donc bien remercier tous les contributeurs pour avoir accepté de présenter les résultats de leurs recherches dans ce format, ainsi que pour leur patience envers les éditeurs de ce volume. Nous remercions également tous les évaluateurs dont les critiques constructives ont été précieuses pour la présentation concise des résultats de recherche dans le cadre des contributions ! Nous tenons à remercier aussi la Fondation German Schweiger (Erlangen) pour son généreux soutien financier, sans lequel la réalisation de ce recueil n'aurait pas été possible. Nous remercions aussi très chaleureusement Dr. Martine Guille (Würzburg) et Baptiste Sellier (Bamberg), dont les conseils nous ont été d'une grande aide. Enfin, nous tenons à remercier les éditeurs des *Romanistische Dossiers* pour intégrer notre livre dans la série ainsi que l'édition AVM Verlag (Munich) pour leur professionnalisme. Merci beaucoup à tous !

Robert Hesselbach und Tanja Prohl
Erlangen/Bamberg, en septembre 2024

Bibliographie

Gabay, Simon; Clérice, Thibault; Reul, Christian. 2023. « OCR17: Ground Truth and Models for 17th c. French Prints (and Hopefully More) ». In : *Journal of Data Mining & Digital* Humanities, 1-14. doi : 10.46298/jdmdh.6492.

Guillot, Céline ; Heiden, Serge ; Lavrentiev, Alexei ; Marchello-Nizia, Christiane. 2008. « Constitution et Exploitation Des Corpus d'ancien et de Moyen Français ». In : *Corpus*. Vol. 7, 1-13. doi : 10.4000/corpus.1495.

Guillot-Barbance, Céline ; Heiden, Serge ; Lavrentiev, Alexei. 2017. « Base de Français Médiéval : Une Base de Référence de Sources