REVUE DE
LINGUISTIQUE
FRANÇAISE
DIACHRONIQUE

2021
28

DIACHRONIQUES

REGARDS LINGUISTIQUES SUR LES ÉDITIONS DE TEXTES MÉDIÉVAUX

Lavretiev, Guillot-Barbance & Heiden – 979-10-231-2172-8

SORBONNE UNIVERSITÉ PRESSES

Regards linguistiques sur les éditions de textes médiévaux

Regards linguistiques sur les éditions de textes médiévaux

SORBONNE UNIVERSITÉ PRESSES
Paris

Les SUP, sont un service général
de la faculté des Lettres de Sorbonne Université.

Diachroniques n° 8

ISBN papier : 979-10-231-0581-0

PDF complet – 979-10-231-2168-1

TIRÉS À PART EN PDF :
Glikman & Verjans – 979-10-231-2169-8
Bragantini-Maillard – 979-10-231-2170-4
Balon – 979-10-231-2171-1
Lavretiev, Guillot-Barbance & Heiden – 979-10-231-2172-8
Mazziotta – 979-10-231-2173-5
Bazin-Tacchella & Souvay – 979-10-231-2174-2

Maquette initiale : Compo-Méca (64990 Mouguerre)
Réalisation : Emmanuel Marc Dubois/3d2s

SUP
Maison de la Recherche
Sorbonne Université
28, rue Serpente
75006 Paris

Tél. (33) 01 53 10 57 60

sup@sorbonne-universite.fr

https://sup.sorbonne-universite.fr

Enjeux philologiques, linguistiques et informatiques de la philologie numérique : l'exemple de la segmentation des mots

Alexei Lavrentiev, Céline Guillot-Barbance & Serge Heiden
Laboratoire IHRIM, ENS de Lyon/CNRS

Depuis les années 1960, le développement des ressources numériques a été particulièrement important en linguistique historique et diachronique, tout spécialement pour la période la plus ancienne du français. Trois types de ressources ont ainsi progressivement vu le jour : des dictionnaires électroniques pour l'ancien et le moyen français (*Dictionnaire du moyen français* [*DMF*], *Anglo-Norman Dictionary*, *Dictionnaire étymologique de l'ancien français* et *Dictionnaire électronique de Chrétien de Troyes*, notamment), des corpus textuels médiévaux (Corpus d'Anthony Dees devenu depuis le « Nouveau corpus d'Amsterdam », base textuelle du *DMF*, base textuelle du Laboratoire de français ancien, Base de français médiéval) et des éditions électroniques de textes phares du Moyen Âge (comme par exemple le *Chevalier de la Charrette*[1], *The Online Froissart*[2], cinq romans de Chrétien de Troyes édités par Pierre Kuntsmann[3], le roman de la *Queste del saint Graal* édité par Christiane Marchello-Nizia et Alexei Lavrentiev[4]).

L'essor de ces ressources textuelles et linguistiques s'est accompagné, dans le même temps, du développement d'outils de recherche et d'analyse de plus en plus élaborés et adaptés à la perspective diachronique. La Base de français médiéval

1. http://www.princeton.edu/~lancelot/ss.
2. http://www.hrionline.ac.uk/onlinefroissart.
3. Ces éditions ont été réalisées dans le cadre du projet de *Dictionnaire électronique de Chrétien de Troyes* (DÉCT, http://www.atilf.fr/dect).
4. http://catalog.bfm-corpus.org/qgraal_cm.

que notre équipe développe à l'ENS de Lyon[5], par exemple, est passée des concordanciers imprimés des années 1990 au logiciel d'interrogation en ligne Weblex dans les années 2000[6], puis, à partir de 2012, à un portail web reposant sur la plateforme TXM[7]. Chaque changement technologique a apporté aux utilisateurs de nombreuses nouvelles fonctionnalités. Le développement conjoint et simultané des ressources textuelles et logicielles a permis la réalisation d'un grand nombre d'études en linguistique diachronique portant aussi bien sur le lexique, que sur la syntaxe, la sémantique grammaticale, la pragmatique, etc.[8].

L'essor de la linguistique diachronique de corpus a également permis d'enrichir l'encodage et l'annotation des textes numériques en fonction des programmes informatiques chargés de les analyser et des questions de recherche qui leur sont posées. Après une phase d'accroissement quantitatif, les corpus et éditions numériques se sont généralement dotés d'outils et de systèmes de catégorisation en vue de l'annotation linguistique des données primaires (lemmes et catégories morphosyntaxiques dans le cas du Nouveau corpus d'Amsterdam[9], catégories morphologiques et structures syntaxiques dans le projet « Les Voies de français »[10] comme dans le corpus Syntactic Reference Corpus of Medieval French[11], lemmes dans les éditions

5. http://txm.bfm-corpus.org.
6. Bénédicte Pincemin, Céline Guillot, Serge Heiden, Alexei Lavrentiev et Christiane Marchello-Nizia, « Usages linguistiques de la textométrie. Analyse qualitative de la consultation de la Base de français médiéval *via* le logiciel Weblex », *Syntaxe & Sémantique*, 9, « Textes, documents numériques, corpus. Pour une science des textes instrumentée », 2008, p. 87-110.
7. Serge Heiden, Jean-Philippe Magué et Bénédicte Pincemin, « TXM : une plateforme logicielle *open source* pour la textométrie – conception et développement », dans Sergio Bolasco, Isabella Chiari et Luca Giuliano (dir.), *Statistical Analysis of Textual Data*, actes de la 10e Journée internationale d'analyse statistique des données textuelles (Rome, juin 2010), Milano, LED, 2010, t. II, p. 1021-1032.
8. Voir, entre autres, Christiane Marchello-Nizia, *L'Évolution du français. Ordre des mots, démonstratifs, accent tonique*, Paris, A. Colin, 1995 ; Sophie Prévost, *La Postposition du sujet en français aux xve et xvie siècles. Analyse sémantico-pragmatique*, Paris, CNRS Éd., 2001 ; Céline Guillot-Barbance, *L'Évolution sémantique du démonstratif en français (ixe-fin xve siècle)*. Deixis*, personne et espace*, Louvain, Peeters, 2017.
9. http://www.uni-stuttgart.de/lingrom/stein/corpus.
10. http://www.voies.uottawa.ca.
11. http://srcmf.org.

de Chrétien de Troyes par Pierre Kunstmann[12] déjà citées, chartes lemmatisées éditées sous la direction de Martin-Dietrich Glessgen[13], corpus lemmatisé de PALM/Meditext[14], etc.).

Cette double activité d'encodage de sources textuelles et d'annotation pour l'extraction et l'analyse d'informations linguistiques basées sur l'indexation systématique de tous les éléments de surface des textes (mots, ponctuations, etc.) amène à poser une nouvelle fois, de façon empirique et sous un jour partiellement nouveau, les questions traditionnellement abordées par l'édition de textes à l'échelle d'un corpus entier : comment, par exemple, traiter la variation morphologique et graphique du texte, sa ponctuation, les délimitations de ses unités linguistiques (mots, locutions, segments de discours comme les passages au discours direct) ? Qu'elles s'appliquent à des ressources créées à partir d'éditions papier numérisées ou à des éditions numériques natives (*born-digital*), les opérations réalisées sur support numérique conduisent toujours à des réflexions méthodologiques et à des choix techniques concernant tous les aspects du texte pris en charge par le travail philologique d'édition. Quelques publications de nature plus méthodologique traitent de ces choix, dont certains seulement sont spécifiques aux états de langue anciens, et de leur exploitation possible pour la recherche linguistique et diachronique[15].

L'accroissement et l'enrichissement continu des ressources numériques a par ailleurs conduit à la mise en place et à la diffusion progressive de bonnes pratiques partagées pour les textes médiévaux (métadonnées attachées à l'unité textuelle, structure

12. Dans le cadre du projet DÉCT ; citées *supra*.
13. http://www.rose.uzh.ch/docling.
14. http://palm.huma-num.fr/PALM.
15. Bénédicte Pincemin, Céline Guillot, Serge Heiden, Alexei Lavrentiev et Christiane Marchello-Nizia, « Usages linguistiques de la textométrie. Analyse qualitative de la consultation de la Base de français médiéval *via* le logiciel Weblex », art. cit. ; Céline Guillot, Alexei Lavrentiev, Bénédicte Pincemin et Serge Heiden, « Le discours direct au Moyen Âge : vers une définition et une méthodologie d'analyse », dans Dominique Lagorgette et Pierre Larrivée (dir.), *Représentations du sens linguistique 5*, Chambéry, Université de Savoie, coll. « Langages », 2013, p. 17-41.

formelle et sémantique interne au texte, annotations diverses)[16]. Ces principes communs permettent le partage, l'échange et l'exploitation des données par un grand nombre d'outils et de chercheurs. Ils rendent également possible l'incrémentation collective, la capitalisation et la pérennisation des traitements et des enrichissements apportés aux données. Par exemple, les textes lemmatisés du projet de *Dictionnaire électronique de Chrétien de Troyes* ont été intégrés à la Base de français médiéval où ils ont bénéficié d'une vérification supplémentaire de l'étiquetage morphosyntaxique, ce qui a permis aux auteurs du dictionnaire, à leur tour, d'améliorer le traitement des mots grammaticaux dans leur base. Ces principes permettent, enfin, la confrontation de pratiques diverses et leur mise en perspective avec les techniques plus anciennes mais à certains égards plus éprouvées de la philologie « traditionnelle ».

Notre article sera centré sur une problématique très précise en rapport avec ces questions générales, celle liée à la délimitation des unités-mots à l'intérieur du texte numérique. Nous essaierons de montrer à partir de cet exemple concret les spécificités et les nouveautés liées à la perspective numérique et leurs implications pour la recherche linguistique, d'une part, et l'édition de textes, d'autre part. Dans un premier temps, nous aborderons les enjeux de la segmentation en mots pour la linguistique diachronique, du point de vue de ses objets et outils de recherche. Nous montrerons dans un second temps quelles sont les difficultés liées au traitement de cette question, puis nous proposerons des solutions adaptées au domaine numérique.

La segmentation du texte en mots : quels enjeux pour la linguistique diachronique de corpus ?

La délimitation des unités lexicales à l'intérieur du texte apparaît comme une question centrale pour la linguistique de corpus, quelles que soient les périodes considérées. Les métho-

16. Voir notamment les recommandations du Consortium international pour les corpus de français médiéval (http://ccfm.ens-lyon.fr).

dologies d'analyse qui caractérisent ce cadre de recherche reposent en effet sur (i) la possibilité de repérer et d'extraire les unités sur lesquelles on travaille, (ii) la possibilité de les dénombrer, (iii) la possibilité de gérer la variation intrinsèque de ces unités, y compris dans leur découpage formel et graphique.

L'apport de l'analyse de l'hétérogénéité des graphies médiévales

Les enjeux de la segmentation en mots, qui peut sembler un problème très pratique, sont en réalité particulièrement importants pour la recherche diachronique. On sait que la délimitation des unités graphiques dans les manuscrits médiévaux diffère dans des proportions très variables de celle du français moderne. La restitution au chercheur d'aujourd'hui des pratiques médiévales authentiques est une donnée importante pour sa connaissance des manuscrits, de leurs conditions matérielles de réalisation, des pratiques et habitudes d'écriture qui se développent et évoluent tout au long de la période médiévale. Elle peut également nous donner de précieux renseignements quant à la conscience linguistique des copistes médiévaux.

Deux types de phénomènes méritent une attention particulière. Les locutions en cours de figement vers un lexème, comme la préposition *parmi*, l'adverbe négatif *jamais*, la conjonction *lorsque*, etc., sont le résultat du processus de grammaticalisation et/ou de lexicalisation de deux unités initialement autonomes. La disparition progressive de l'espace blanc entre les éléments de ces locutions dans les pratiques scribales peut servir d'indice précieux pour l'analyse du processus de grammaticalisation/lexicalisation. Le phénomène inverse, qui tend à dissocier linguistiquement et graphiquement deux unités antérieurement soudées, est beaucoup moins fréquent, mais tout aussi intéressant. L'étude diachronique de l'adverbe *très* du français moderne[17], par exemple, a montré l'importance de cette étape graphique dans le passage du préverbe ou de la

17. Christiane Marchello-Nizia, *Grammaticalisation et changement linguistique*, Bruxelles, De Boeck/Duculot, 2006, p. 166-172.

particule *tres*, toujours accolée au mot qui la suit, au statut d'adverbe autonome. Les habitudes graphiques et leurs changements progressifs entrent dans ce cas en relation directe avec les évolutions linguistiques et montrent l'importance pour le linguiste de pouvoir prendre en compte la réalité matérielle du découpage lexical du texte médiéval. Les deux phénomènes que nous venons de citer sont à distinguer nettement d'autres habitudes graphiques médiévales, sans lien direct avec ces processus de formation d'unités lexicales, comme la tendance à souder les unités grammaticales atones (pronoms personnels, articles, etc.) au mot plein qui les précède ou qui les suit. Mais, même dans ces derniers cas apparemment moins importants pour l'approche diachronique, la segmentation médiévale n'est pas sans intérêt pour le linguiste (elle n'est pas sans liens avec les phénomènes d'élision ou d'enclise, ou d'absence d'autonomie de mots grammaticaux).

Nous proposons d'adopter une terminologie précise pour la désignation de ces différents phénomènes : nous utiliserons le terme *agglutination* pour désigner la pratique scribale qui consiste à souder graphiquement deux unités lexicales ou les composants d'une locution figée en cours de lexicalisation et le terme *déglutination* pour la pratique inverse qui consiste à séparer graphiquement les parties d'une seule unité lexicale ou d'une locution figée.

L'apport de l'usage de graphies normalisées

La nécessité de rendre compte de la matérialité du texte médiéval, dans son caractère instable et évolutif, se heurte à un principe concurrent et antagoniste : la nécessité de devoir repérer et compter les unités lexicales sur lesquelles on travaille oblige à dépasser ou à éliminer les variations formelles qui les opposent. Ces variations sont de deux ordres : soit elles résultent de traits caractéristiques qui distinguent les manuscrits de types, de régions ou d'époques différents, soit ces variations sont internes à un même document manuscrit et ne font qu'illustrer la tendance forte de la culture médiévale à toujours varier sur le même motif.

Par ailleurs, les annotations et les enrichissements linguistiques dont il a été question plus haut portent, pour un grand nombre d'entre eux (lemmatisation, étiquetage morphosyntaxique, annotation sémantique), sur l'unité linguistique fondamentale qu'est le mot. L'association de son lemme ou de sa catégorie morphosyntaxique à une unité lexicale présuppose que ses limites formelles aient été déterminées au préalable et qu'elles l'aient été de manière suffisamment cohérente et systématique pour permettre une bonne exploitation de cette information linguistique. La normalisation la plus systématique possible est par conséquent souhaitable pour faciliter l'annotation et l'interrogation de corpus. Elle l'est également pour améliorer la performance des outils de traitement automatique de la langue (TAL), tels que les étiqueteurs morphosyntaxiques et les lemmatiseurs.

Ces deux principes divergents, à savoir le respect du manuscrit dans sa variation interne et externe et la normalisation de cette variation en vue de recherches portant non pas sur des formes mais sur des catégories linguistiques, s'appliquent en réalité à différents niveaux du texte ou à différents usages de la représentation numérique : les outils de visualisation doivent permettre de lire le texte tel qu'il se présente dans ses réalisations médiévales effectives ; les outils de requête et d'annotation doivent permettre de regrouper, comparer et analyser le fonctionnement d'unités linguistiques en contexte[18]. Nous proposerons à la fin de cette contribution quelques solutions pratiques qui répondent à ces usages multiples.

18. Voir notamment Nicolas Maziotta, « Le texte dans tous ses états. Philosophie d'encodage du projet Khartês », dans Gérald Purnelle, Cédrick Fairon et Anne Dister (dir.), *Le Poids des mots*, actes des 7e Journées internationales d'analyse statistique des données textuelles, Louvain-la-Neuve, UCL/Presses universitaires de Louvain, 2004, t. II, p. 793-803 ; Alexei Lavrentiev, *Tendances de la ponctuation dans les manuscrits et incunables français en prose, du XIIIe au XVe siècle* [thèse de doctorat en sciences du langage soutenue sous la dir. de Christiane Marchello-Nizia, ENS Lyon, 2009], en ligne.

La segmentation du texte en mots : difficultés théoriques et pratiques

Les difficultés théoriques et pratiques engendrées par la segmentation du texte en unités-mots se manifestent du côté de la philologie traditionnelle tout autant que des ressources numériques.

Les pratiques de segmentation de la philologie traditionnelle

On constate que la question de la segmentation en mots est très peu traitée dans les manuels consacrés à l'édition de textes et qu'elle n'est souvent abordée que rapidement, voire pas du tout, dans les introductions linguistiques aux éditions papier[19]. Ainsi, le rapport de la commission des romanistes chargée de l'établissement des règles pratiques pour l'édition des anciens textes français et picards[20] n'en porte aucune mention. Tel est aussi le cas du *Guide de l'édition de textes en ancien français* paru trois quarts de siècle plus tard[21]. Alfred Foulet et Mary Blakely Speer[22] consacrent quelques pages aux problèmes des mots composés et de la réduplication des consonnes finales ou initiales en cas de soudure graphique (*derrechief < de rechief*, *dessus < de sus*, etc.). Ils notent que la frontière entre une locution figée et un mot composé peut évoluer dans le temps, et prennent comme exemple l'expression *avoir a faire a quelqu'un*, devenue plus tard *avoir affaire à quelqu'un*. Ils ne donnent cependant pas de règles ou d'indices pour définir le moment où ce changement de segmentation a lieu. Ils insistent seulement sur la nécessité d'être attentif aux cas où le choix de la segmentation influe

19. La situation est différente dans le domaine de l'édition diplomatique des chartes où l'attention aux aspects graphiques est plus importante dès les années 1970 (voir *Documents linguistiques de la France.* éd. dir. Jean-Gabriel Gigot, Jacques Monfrin et Lucie Fossier, Série française I, *Chartes en langue française antérieures à 1271 conservées dans le département de la Haute-Marne*, Paris, Éd. du CNRS, 1974).
20. Mario Roques, « Établissement des règles pratiques pour l'édition des anciens textes français et provençaux », *Romania*, 52, 1926, p. 243-249.
21. Yvan G. Lepage, *Guide de l'édition de textes en ancien français*, Paris, Champion, 2001.
22. Alfred Foulet et Mary Blakely Speer, *On Editing Old French Texts*, Lawrence, Regents Press of Kansas, 1979, p. 60-62.

sur le sens de la locution. Ainsi, la préposition *devers* indique la direction d'un mouvement, tandis que la locution *de vers* indique la provenance. Dans les cas de « désambiguïsation par segmentation », ils n'hésitent pas à recommander de supprimer les consonnes réduplicuées, comme lorsque *dessus* a le sens de « de dessus » et doit être transcrit *de sus*. Il convient de noter que cette différence sémantique n'est pas toujours évidente, dans les cas où le verbe ne dénote pas un mouvement orienté. Les éditeurs semblent dans la pratique privilégier la graphie soudée (97 % et 99,5 % respectivement des occurrences de *avoir affaire* et *dessus* dans la Base de français médiéval). Dans l'un des rares textes où l'on trouve les deux graphies, les éditeurs appliquent probablement la règle de Foulet-Speer, même si on peut se demander si la distinction créée n'est pas artificielle, comme dans les deux exemples suivants :

(1) Vit Achelor a la fenestre, / jouste un pilier **de vers** senestre[23].

(2) Si com l'aventure ert a estre, /ambedui joustent **devers** destre[24] ;

Les articles de recherche sur les pratiques scribales de segmentation des mots[25] ne semblent pas avoir influencé considérablement les habitudes des philologues dans les années 1990 et 2000.

Les auteurs des *Conseils pour l'édition des textes médiévaux*[26] recommandent de souder les « expressions composées usuelles ou celles qui sont passées soudées dans la langue moderne », en tenant compte d'une part des graphies (par exemple, distinguer *desorenavant* de *des ore en avant*) et, d'autre part,

23. *Le Roman de Thèbes*, éd. Guy Raynaud de Lage, Paris, Champion, coll. « Classiques français du Moyen Âge », 1966, v. 3510.

24. *Ibid*., v. 5716.

25. Peter Rickard, « Système ou arbitraire ? Quelques réflexions sur la soudure des mots dans les manuscrits français du Moyen Âge », *Romania*, vol. 103, n° 412, 1982, p. 470-512 ; Nelly Andrieux-Reix et Simone Monsonégo, « Écrire les phrases au Moyen Âge. Matériaux et premières réflexions pour une étude des segments graphiques observés dans des manuscrits français médiévaux », *Romania*, vol. 115, n° 459-460, 1997, p. 289-336.

26. Françoise Vielliard et Olivier Guyotjeannin, *Conseils pour l'édition des textes médiévaux* [2001-2002], 3 vol., nouv. éd. revue et mise à jour, Paris, École nationale des chartes/ CTHS, 2014-2018.

de la « réalité institutionnelle » (comme dans *lieuxtenens* pour « lieutenants du roi »). En cas de lettres redoublées, ils proposent de séparer les mots en conservant la double consonne (par exemple *a sses parenz*). Cette dernière recommandation s'appuie notamment sur la pratique de Philippe Ménard dans son édition du *Roman de Tristan en prose*[27], mais semble être très peu suivie dans l'ensemble.

On peut conclure que la règle plus ou moins implicite qui ressort de ces recommandations est d'adapter les pratiques de division aux normes du français moderne pour faciliter la lecture et la compréhension du texte par le lecteur contemporain. Mais sur ce point comme bien d'autres, une certaine latitude est laissée au philologue, dont les choix pourront varier en fonction du texte et du manuscrit qu'il édite, en fonction du public visé et, finalement, en fonction de sa libre appréciation personnelle. Les règles communes n'étant ni très stables ni très explicites, les éditeurs ne mentionnent que rarement les problèmes de segmentation dans leurs introductions linguistiques.

Parmi les rares exceptions, on peut citer Philippe Ménard, qui, dans l'introduction à son édition du *Roman de Tristan en prose*[28], explicite les principes qu'il a retenus pour la segmentation des mots et tient compte, dans certains cas, de l'usage du copiste. Plus récemment, Nelly Andrieux-Reix a consacré une section entière à la « syntaxe graphique » dans l'introduction à son édition du *Moniage Guillaume*[29] et a utilisé les traits d'union et les tirets longs pour représenter les usages du copiste (le trait d'union pour les agglutinations et le tiret ordinaire long pour les déglutinations). Cette pratique n'a pas été suivie par les autres philologues, à l'exception de Frédéric Duval, qui, dans son édition du *Dit de la fleur de lis* de Guillaume

27. *Le Roman de Tristan en prose*, éd. dir. Philippe Ménard, t. I, *Des aventures de Lancelot à la fin de la « Folie Tristan »*, Genève, Droz, coll. « Textes littéraires français », 1987.
28. *Ibid*., p. 53-54.
29. *Le Moniage Guillaume, chanson de geste du XII^e siècle*, éd. de la rédaction longue par Nelly Andrieux-Reix, Paris, Champion, coll. « Classiques français du Moyen Âge », 2003.

de Digulleville[30], distingue l'édition critique de l'édition des témoins, dans laquelle il utilise le tiret bas pour les agglutinations et le tiret pour les déglutinations. Par ailleurs, dans son édition commentée d'un extrait du *Pèlerinage de l'âme* (toujours de Guillaume de Digulleville)[31], le même Frédéric Duval consacre toute une section de son introduction aux problèmes liés à la séparation des mots. Il y explique les pratiques du copiste et ses propres choix d'édition. Enfin, les problèmes de segmentation lexicale occupent une part importante de l'introduction apportée à l'édition numérique de la *Queste del saint Graal*[32]. L'expérience de cette édition a largement inspiré la solution philologique présentée ci-dessous.

Ces quelques exceptions mises à part, il est assez difficile de déterminer quels sont les motifs qui ont conduit l'éditeur à faire ses choix dans le texte qu'il édite. Cette situation engendre une grande diversité de pratiques, en particulier pour les expressions en cours de figement ou en voie d'autonomisation linguistique dont il a été question plus haut. La soudure, non signifiante, des mots grammaticaux (pronoms, articles et prépositions, etc.) n'est à peu près jamais représentée dans les éditions de textes du Moyen Âge, et, pour le reste, les usages ne sont pas très stables (voir la section suivante). La seconde conséquence de cette situation est qu'elle ne permet pas d'apprécier si les segmentations choisies par l'éditeur résultent d'une volonté de normalisation du texte ou du respect des solutions graphiques du manuscrit de base. On note, toutefois, que l'une des tendances les plus répandues est de choisir une segmentation homogène dans un texte donné (l'éditeur écrira de manière systématique *par mi* ou *parmi*), ce qui élimine la variation graphique interne

30. Guillaume de Digulleville, *Le Dit de la fleur de lis*, éd. Frédéric Duval, Paris, École nationale des chartes, coll. « Mémoires et documents de l'École des chartes », 2014.

31. Frédéric Duval, *Descente aux enfers avec Guillaume de Digulleville. Édition et traduction commentées d'un extrait du* Pèlerinage de l'âme *(Paris, BnF, fr. 12466)*, Saint-Lô, Archives départementales de la Manche, 2006.

32. Christiane Marchello-Nizia, Alexei Lavrentiev et Céline Guillot, « Édition électronique de la *Queste del saint Graal* », dans David Trotter (dir.), *Manuel de la philologie de l'édition*, Berlin/Boston, De Gruyter, 2015, p. 155-176.

à chaque manuscrit, permet une certaine harmonisation du texte et en facilite l'appropriation par le lecteur d'aujourd'hui.

Les pratiques de segmentation des textes numériques

Les corpus textuels étant généralement composés d'éditions papier numérisées, ils héritent des pratiques divergentes et souvent peu documentées des éditeurs de textes.

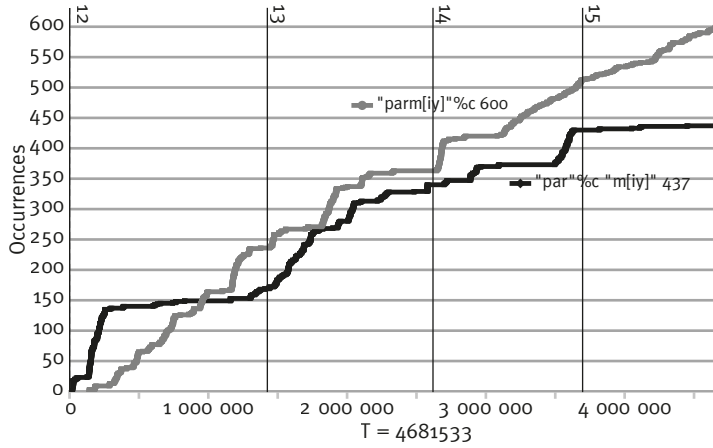
À titre d'exemple, nous nous sommes intéressés à quatre locutions dont la segmentation a évolué dans l'histoire du français : la préposition *par* suivie du substantif *mi* qui a abouti à la préposition *parmi*, le groupe d'adverbes *ja mais* qui a formé *jamais*, le préverbe *en* suivi du verbe *porter* qui a produit, dans certaines constructions, le verbe *emporter,* et enfin le groupe *à* + *venir* qu'il faut relier au verbe *advenir* et au substantif *avenir* et qui continue d'exister en tant que locution adjectivale (au sens de « futur »).

Nous avons recensé les occurrences des différentes formes graphiques (séparées et soudées) de ces locutions dans le corpus BFM2016 de la Base de français médiéval. Le moteur de recherche CQP[33] intégré à la plateforme TXM donnant accès aux textes permet en effet de sélectionner plusieurs variantes graphiques dans une seule expression de requête CQL[34]. L'outil Progression (actuellement disponible dans la version bureau de TXM seulement) sert à visualiser la densité d'apparition d'un motif recherché au sein du corpus, tandis que l'outil Concordance présente l'ensemble des occurrences dans leur contexte en vue d'une analyse plus fine[35].

En ce qui concerne la locution *par* + *mi*, le graphique de progression (**fig. 1**) montre que les deux graphies, soudée (correspondant à la courbe cumulative des occurrences de l'expression de requête `"parm[iy]"%c`[36]) et séparée (corres-

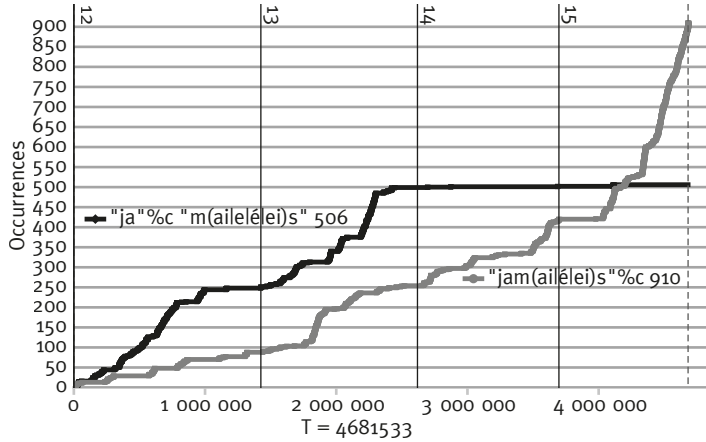
33. http://cwb.sourceforge.net.
34. http://cwb.sourceforge.net/files/CQP_Tutorial.
35. Nous remercions Bénédicte Pincemin (IHRIM), qui nous a apporté son aide pour l'interprétation des graphiques qui suivent.
36. Cette requête permet de retrouver les formes qui commencent par *par* et finissent par *i* ou par *y*. L'opérateur « `%c` » permet d'ignorer la casse.

pondant à la courbe de l'expression de requête `"par"%c "m[iy]"`), sont utilisées avec une fréquence comparable dans les éditions de textes composés entre la fin du XII^e^ et la fin du XIV^e^ siècle (leurs pentes sont comparables, sachant que les apparitions des locutions au sein des textes sont représentées par des marches dans la courbe). La graphie soudée n'est pas utilisée dans les éditions des textes les plus anciens (avant 1150) et la graphie séparée disparaît presque entièrement dans les éditions des textes du XV^e^ siècle. Les courbes comportent cependant de nombreuses alternances de montées/plateaux, qui semblent manifester l'impact fort, sinon déterminant des choix éditoriaux (il y a les textes/éditions qui suivent telle segmentation, et les textes/éditions qui suivent telle autre).



1. Graphique de progression des occurrences des graphies soudées et séparées de *parmi*/*par mi* dans le corpus BFM2016.

Dans le cas de *ja* + *mais* (**fig. 2**), la graphie séparée (correspondant à l'expression de requête `"ja"%c "m(ai|e|é|ei)s"`) est préférée par les éditeurs des textes du XII^e^ siècle. Pour les textes du XIII^e^ siècle, l'usage varie selon les éditeurs, et la graphie soudée (correspondant à l'expression de requête `"jam(ai|é|ei)s"%c`) devient pratiquement exclusive pour les éditions des textes composés à partir de la fin du siècle.

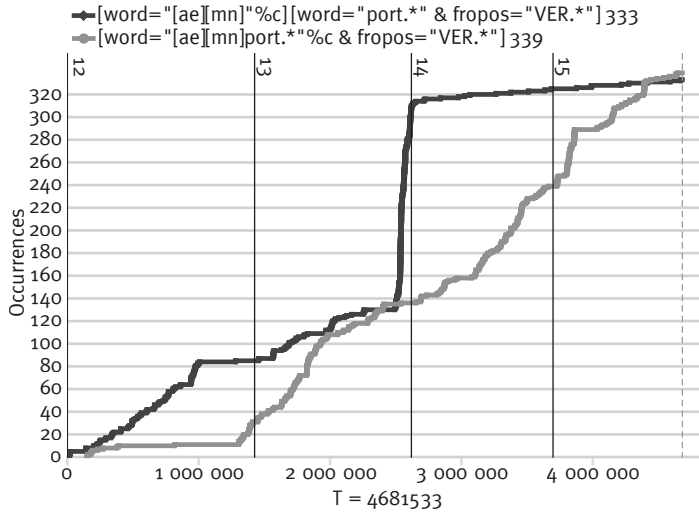


2. Graphique de progression des occurrences des graphies soudées et séparées de *jamais*/*ja mais* dans le corpus BFM2016.

En ce qui concerne le verbe *en porter*/*emporter*, le graphique de progression (**fig. 3**) est particulièrement intéressant, car on observe un « saut » énorme de la graphie séparée (correspondant à l'expression de requête `[word="[ae][mn]"%c] [word="port.*" & fropos="VER.*"]`) à la fin du XIII^e^ siècle, tandis que la graphie soudée (correspondant à l'expression de requête `[word="[ae][mn]port.*"%c & fropos="VER.*"]`) progresse d'une manière à peu près constante à partir de la fin du XII^e^ siècle. Le « saut » de la fin du XIII^e^ siècle s'explique par la fréquence exceptionnellement élevée de ce verbe dans les *Coutumes de Beauvaisis* de Philippe de Beaumanoir (181 occurrences sur 672 dans toute la BFM2016, toutes graphies confondues), toujours graphié séparément par Amédée Salmon[37]. Notons qu'une grande partie des rares occurrences de la graphie séparée attestées dans les éditions de textes à partir du XIV^e^ siècle provient des constructions où *en* joue clairement le rôle d'un pronom et où la graphie soudée ne serait pas correcte du point de vue de l'orthographe moderne :

37. À l'exception d'une occurrence en bas de la p. 238, où il peut s'agir d'une coquille.

(3) Vous estes chevalier ? dist le Jouvencel, car vous **en portés** les enseignes, ce me semble[38].



3. Graphique de progression des occurrences des graphies soudées et séparées de *emporter/en porter* dans le corpus BFM2016.

Le cas de la locution *à venir/a(d)venir* est plus complexe (correspondant respectivement aux expressions de requête `"ad?venir"%c | "[aà]d?"%c "venir"`). Le graphique de progression (**fig. 4**) montre que les deux formes sont utilisées tout au long de la période médiévale, avec une prépondérance constante de la graphie soudée (sauf dans les plus anciens textes du XII^e^ siècle). Cette persistance de la double segmentation s'explique certainement par la diversité des fonctions de cette locution, qui a abouti à trois formes distinctes dans l'orthographe moderne : *à venir*, *avenir* et *advenir*. Dans les usages au sens de « survenir », la graphie soudée semble la seule possible dès les premières attestations :

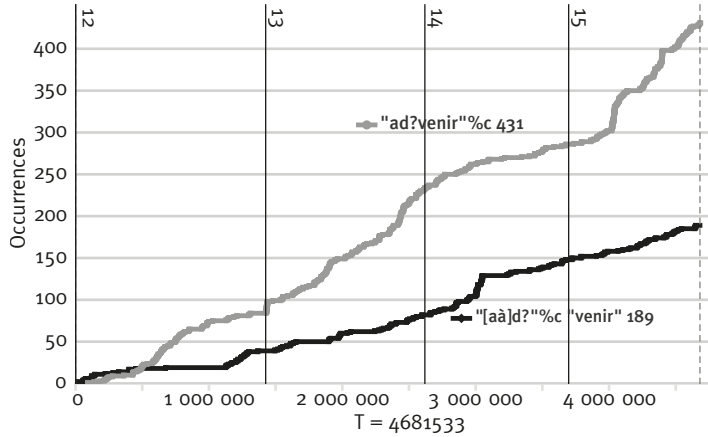
38. Jean de Bueil, *Le Jouvencel*, éd. Léon Lecestre, comm. Guillaume Tringant, Paris, H. Laurens, t. I, 1887, p. 99.

(4) Si m'ait Deus, qui ne menti, / jeo nel lerroie pur murir, / que jeo ne l'auge ja ferir, / que ke m'en deie **avenir**[39].

Dans les usages où la préposition *à* est régime d'un verbe conjugué, la graphie séparée s'impose à son tour :

(5) Ci comencent angoisses dolentes **a venir**[40].

En revanche, dans les constructions *estre a venir* et *temps a venir*, on trouve aussi bien les graphies soudées que séparées.



4. Graphique de progression des occurrences des graphies soudées et séparées de *avenir*/*a venir* dans le corpus BFM2016.

Comme on l'a indiqué plus haut, les éditeurs sont généralement constants dans le choix de la graphie au sein d'un texte donné, même si des exceptions sont possibles.

Toutes ces observations nous permettent de conclure qu'il y a une tendance générale chez les éditeurs à préférer les graphies segmentées pour les textes les plus anciens (avant la fin du XII^e^ siècle) et que les graphies soudées dominent largement à partir du XIV^e^ ou du XV^e^ siècle, selon les cas. Pour les textes composés au XIII^e^ siècle, les pratiques des éditeurs sont divergentes.

39. *Gormont et Isembart. Fragment de chanson de geste du XII^e^ siècle* [*ca* 1130], éd. Alphonse Bayot, Paris, Champion, coll. « Classiques français du Moyen Âge », 3^e^ éd., 1931, v. 211.
40. *Li ver del juïse. Sermon en vers du XII^e^ siècle* [2^e^ quart du XII^e^ siècle], éd. Erik Rankka, Stockholm, Almqvist och Wiksell, 1982, v. 276.

L'hésitation dans les choix de segmentation se répercute également dans les dictionnaires. Dans le *Dictionnaire de moyen français* par exemple, on trouve des entrées pour les mots soudés, mais aussi des exemples sous les entrées du dictionnaire correspondant à l'une des parties de la locution. Ainsi des exemples de la forme *advenir* (au sens adjectival de « à venir ») peuvent être retrouvés sous les entrées « advenir » (III.A), « avenir » (I) et « venir » (II.A.4.a), et il n'y a pas de renvois systématiques d'un lemme à l'autre.

La segmentation du texte en mots : méthode de la philologie numérique

À ces variations du niveau graphique répond la tendance à l'harmonisation systématique des traitements numériques. La nécessité de normaliser les différences formelles à l'intérieur d'un même corpus textuel transpose le problème du niveau scribal au niveau éditorial : ce ne sont plus les variations constantes des copistes médiévaux qu'il s'agit de dépasser, mais les pratiques hétérogènes des éditeurs de textes, eux-mêmes confrontés à une documentation écrite aux règles instables et en constante évolution.

D'un autre côté, si les pratiques des éditeurs ne fournissent généralement pas de données fiables pour l'étude de la segmentation graphique dans les manuscrits médiévaux, elles constituent en même temps une donnée non négligeable de l'histoire de la pensée philologique. À ce titre, elles méritent d'être observées et conservées. Pour dépasser ces points de vue apparemment antagonistes, on peut souligner que la normalisation de la représentation numérique n'empêche pas l'encodage supplémentaire des pratiques des scribes et des éditeurs scientifiques. Le défi de la philologie numérique consiste ainsi à proposer des solutions d'encodage qui permettent de produire des éditions et des corpus numériques à la fois normalisés, faciles à utiliser et conservant les données primaires (pratiques des scribes) et secondaires (pratiques des éditeurs) potentiellement utiles.

Pour que la machine soit en mesure de relever ce défi, il faut modéliser informatiquement la segmentation des mots. Cette méthode repose sur trois opérations distinctes :

1. la représentation philologique numérique explicite de la séparation ou de la soudure lexicale ;
2. l'interprétation correcte de cette représentation par les moteurs de recherche ;
3. la possibilité de générer les visualisations de niveaux linguistiques nécessaires à partir de cette représentation.

Il est nécessaire de distinguer le versant philologique de la méthode qui consiste à définir les principes de normalisation et de restitution des données primaires ou secondaires non normalisées, et la solution numérique de la méthode qui permet d'implémenter les choix philologiques aux niveaux de l'encodage, de l'interrogation et de la visualisation des ressources.

Versant philologique de la méthode

Comme nous l'avons vu plus haut, les problèmes de segmentation lexicale ont commencé à attirer l'attention de certains éditeurs scientifiques à la fin des années 1980 et surtout au cours des années 2000. L'édition numérique de la *Queste del saint Graal*[41], dont les débuts remontent à 1999, a été l'occasion de réviser la méthodologie traditionnelle d'établissement du texte afin de profiter des avantages offerts par le numérique et de créer une ressource d'un nouveau type, prototype pour de futurs projets éditoriaux.

Cette édition électronique présente la particularité d'offrir à l'utilisateur différents niveaux de transcription et de lecture du texte : (i) un « niveau normalisé », qui facilite la lecture en adoptant un mode de présentation plus conforme aux usages modernes ; (ii) un «niveau diplomatique », plus respectueux de la source primaire et signalant les interprétations de l'éditeur ; (iii) un « niveau fac-similaire », qui reproduit de manière

41. Éd. Christiane Marchello-Nizia et Alexis Lavrentiev, Lyon, ENS Lyon, en ligne : http://catalog.bfm-corpus/org/qgraal_cm.

aussi fidèle que possible le document médiéval dans ses particularités graphiques et matérielles[42]. Ces trois niveaux de lecture (ou « facettes ») sont bien distincts du niveau de base (ou « de référence ») correspondant aux unités linguistiques retenues pour les requêtes de recherches et de décomptes et l'étiquetage morphosyntaxique du texte. Et l'interface de la plateforme TXM permet de coupler l'affichage simultané d'un ou de plusieurs niveaux avec le moteur de recherche CQP[43].

Le traitement des locutions en cours de figement a fait l'objet d'une attention particulière dans cette édition pilote. L'introduction[44] explicite les règles de transcription retenues, mais elle présente aussi un tableau exhaustif des locutions en cours de figement avec leur traitement par le scribe (graphie soudée, séparée par un espace ou par un passage à la ligne) et les graphies retenues (indexées) pour les requêtes et l'étiquetage morphosyntaxique. Les graphies du scribe sont toujours préservées à l'affichage pour ces locutions, y compris dans la version normalisée du texte. L'expression de requête pour la forme est choisie au cas par cas et correspond le plus souvent à la graphie majoritaire du scribe. C'est par exemple le cas de la locution conjonctive *puis que* écrite avec une espace dans 74 occurrences sur 84, ou de l'adverbe *jamais* avec 78 graphies soudées sur 83. En revanche, la forme soudée a été retenue pour les requêtes portant sur la préposition *jusque*, bien que le manuscrit présente la graphie séparée *jus que* dans 72 occurrences sur 85. Dans le cas de l'adverbe ou locution adverbiale *a-tant*, la variation graphique du manuscrit (25 occurrences soudées contre 19 séparées) a été préservée dans les formes indexées pour les requêtes. Ces choix, parfois hésitants, témoignent de l'évolution de la réflexion méthodologique au cours du projet éditorial.

42. Christiane Marchello-Nizia, Alexei Lavrentiev et Céline Guillot, « Édition électronique de la *Queste del saint Graal* », art. cit.
43. Éd. cit. (voir *supra*, n. 40).
44. Base de français médiéval, introduction, p. 32-36.

Les douze premiers paragraphes (soit près de deux feuillets) du *Graal* ont été transcrits au niveau « fac-similaire » (ou « allographétique ») en plus des deux niveaux, « normalisé » et « diplomatique », disponibles pour l'ensemble du texte. Le niveau facsimilaire tient compte des variantes de lettres (ou « allographes », comme les *s* « rond » et « long ») et des marques d'abréviation, mais aussi des segmentations particulières (agglutinations et déglutinations), qu'elles soient ou non en cours de figement. On peut par exemple observer des graphies soudées *fuuenu* (pour « fu venu ») et *ſigrant* (pour « si grant ») à la ligne 2 de la colonne 160a du manuscrit K. Le statut linguistique de ces agglutinations étant *a priori* différent de celui des locutions en cours de figement (elles ne sont pas liées à des phénomènes de grammaticalisation ou de lexicalisation), il est important de pouvoir distinguer les deux types d'agglutination tant au niveau de la visualisation qu'au niveau des requêtes. Les agglutinations sans figement ne sont jamais prises en compte dans l'indexation des mots du texte, et ne sont pas visibles aux niveaux normalisé et diplomatique de la transcription. En revanche, on peut rechercher les occurrences correspondantes grâce à une annotation spéciale (expression CQL `[rend="aggl"]`) et les visualiser dans la transcription fac-similaire.

Les sources XML-TEI de l'édition électronique de la *Queste del saint Graal* ont été distribuées et utilisées dans divers projets de recherche. Elles ont surtout été enrichies d'un niveau d'annotation supplémentaire dans le cadre du projet franco-allemand « Syntactic Reference Corpus for Medieval French »[45]. L'annotation syntaxique basée sur les unités lexicales a permis de valider, et dans certains cas de réviser, la segmentation des unités d'après l'analyse syntaxique du texte. Et l'annotation parallèle de plusieurs autres textes médiévaux a permis de confronter nos choix à un corpus diachronique d'œuvres

45. Achim Stein et Sophie Prévost, « Syntactic annotation of medieval texts: the Syntactic Reference Corpus of Medieval French (SRCMF) », dans Paul Bennett, Martin Durrell, Silke Scheible et Richard J. Whitt (dir.), *New Methods in Historical Corpora*, Tübingen, Narr, coll. « Corpus Linguistics and International Perspectives on Language (CLIP) », 2013, t. III, p. 275-282 ; voir, en ligne : http://srcmf.org.

composées entre le IXe et la fin du XIIIe siècle. Des principes de segmentation plus simples, constants et généraux ont été ainsi définis. Leur atout principal est d'être applicables à n'importe quel texte de français médiéval.

Ces principes reposent sur une règle fondamentale très simple : l'indexation des unités lexicales (ou *tokenisation*) se base systématiquement sur la segmentation maximale (la plus fine possible). Des solutions informatiques sont proposées pour (i) afficher la segmentation du manuscrit de base et/ou la segmentation décidée par l'éditeur scientifique, et (ii) faciliter les requêtes sur les locutions en cours de figement, qu'elles soient graphiées avec ou sans l'espace blanc.

Dans la pratique, l'application de ce principe amène parfois à des coupures contre-intuitives, comme dans le cas de *bon heur*, mais elle présente l'avantage d'offrir davantage de souplesse pour l'indexation et l'annotation du texte, car il est toujours plus facile de regrouper des éléments annotés que de créer des annotations sur des parties d'éléments indivisibles. La seule exception à cette « atomisation » de la segmentation concerne les cas où la fonction de l'un au moins des éléments séparables est très difficile à définir en contexte et à catégoriser à l'aide d'une étiquette morphosyntaxique. C'est par exemple le cas de *que* final dans *quelque* (même si des graphies séparées sont attestées dans les manuscrits).

Ce principe de segmentation est appliqué systématiquement dans les nouvelles éditions de la collection « Sources médiévales » mise en place à l'École normale supérieure de Lyon et liée à la Base de français médiéval. Progressivement, le *Graal* et l'ensemble des textes de la base seront ré-indexés en conformité avec ces nouvelles normes.

Versant numérique de la méthode

La méthode informatique utilisée pour traiter la segmentation lexicale comporte trois aspects :

1. l'encodage des données, qui établit une relation contractuelle avec les outils à travers un format associé à des conventions d'interprétation ;
2. l'ergonomie de la saisie, de la vérification et de l'enrichissement des données ;
3. l'ergonomie de l'exploitation des données (recherche, lecture et exportation des résultats et ré-annotation).

Les technologies d'encodage de données textuelles et paratextuelles sont extrêmement diverses et ont tendance à évoluer rapidement. Afin d'assurer une certaine pérennité à la solution adoptée, il est indispensable d'utiliser des standards internationaux soutenus par des communautés scientifiques importantes et ayant fait la preuve de leur capacité d'adaptation aux changements technologiques. Dans le domaine de l'édition numérique académique, les recommandations du consortium *Text Encoding Initiative*[46] répondent parfaitement à ces exigences. Mais dans le domaine de la linguistique de corpus et du traitement automatique du langage naturel (TAL), la TEI est loin de s'imposer. Néanmoins, le cadre proposé par la TEI (un jeu de près de 700 balises au format XML[47], bien documenté et extensible) constitue certainement l'une des meilleures options lorsqu'on veut combiner la richesse du balisage philologique avec la puissance des outils de TAL et d'analyse de corpus. C'est pourquoi la Base de français médiéval et la plateforme TXM ont choisi une représentation XML-TEI des textes numériques comme base commune d'interopérabilité.

L'élément <w> de la TEI[48] sert à encoder un mot ou une unité de segmentation lexicale (*token*). Cette unité peut porter de multiples annotations, dont peut faire partie la segmentation graphique primaire ou secondaire. La TEI ne spécifie pas quelle forme doivent prendre ces types particuliers d'annotation, mais une convention plus précise peut être adoptée dans le cadre d'un

46. TEI, http://www.tei-c.org.
47. https://www.w3.org/TR/2006/REC-xml11-20060816.
48. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-w.html. Les références de toutes les autres balises de la TEI peuvent être consultées de manière analogue.

projet ou d'une communauté spécifique. Dans le cadre de la Base de français médiéval et de la collection d'éditions numériques associée, on a fait le choix d'indiquer l'absence « anormale » d'un espace après un *token*. La locution prépositionnelle *par* + *mi* soudée graphiquement dans la source sera alors encodée comme ceci :

```
<seg type="fgmt">
  <w rend="aggl">par</par>
  <w>mi</w>
</seg>
```

L'élément <seg> qui englobe l'ensemble est utilisé pour distinguer les agglutinations dans les locutions en cours de figement (signalées par la valeur « fgmt » de son attribut *type*) des autres cas (moins réguliers et *a priori* moins intéressants du point de vue linguistique). Dans les rares cas où il est nécessaire de noter la présence d'une espace à l'intérieur d'un seul *token*, nous avons fait le choix d'utiliser la balise <space/> :

```
<w>quel<space quantity="1" unit="chars"/>que</w>
```

Il s'agit en effet d'un cas exceptionnel, puisque le principe général consiste à segmenter le plus possible.

Bien entendu, les balises <w> et leurs annotations n'ont pas vocation à être encodées « à la main » par l'éditeur du texte. La « tokenisation » est réalisée par des outils de TAL plus ou moins sophistiqués et une méthode ergonomique peut être mise en place pour pré-annoter les cas particuliers, comme les agglutinations, et pour contrôler le résultat de l'application de ces outils[49]. Dans la collection « Sources médiévales », par exemple, le caractère de pré-annotation « + » sera utilisé après le premier élément d'une agglutination simple (*des+ coses* : « des choses »), et un double plus « ++ » marque les agglutinations dans les locutions en cours de figement (*par++ mi*). Le tiret bas est utilisé dans les cas de déglutinations (*a_vons* : « avons »). Il est dupliqué dans les mots composés (*quel__que*). Dans certains

49. Voir Nicolas Maziotta, « Le texte dans tous ses états. Philosophie d'encodage du projet Khartês », art. cit. ; Alexei Lavrentiev, *Tendances de la ponctuation dans les manuscrits et incunables français en prose, du XIII^e^ au XV^e^ siècle*, thèse cit.

cas, la lecture du manuscrit ne permet pas de déterminer avec certitude si un espace blanc est présent entre deux mots ou entre deux parties d'un même mot. Il est alors possible d'encoder cette incertitude en ajoutant un point d'interrogation à la suite du caractère « + » ou « _ ».

Les caractères de pré-annotation sont interprétés par le logiciel de segmentation lexicale (ou « tokeniseur ») qui produit les balises TEI mentionnées plus haut. L'ensemble des règles de saisie de transcription des sources est présenté dans un document en ligne, constamment mis à jour[50].

Le balisage explicite des mots du corpus avec l'annotation des segmentations particulières permet d'obtenir tous les affichages souhaités pour la lecture du texte : on peut, au choix, imiter la segmentation graphique de la source ou, au contraire, « normaliser » tout ou partie des segmentations particulières. Dans le cas de la collection « Sources médiévales », le choix a été fait de conserver la segmentation graphique des manuscrits pour les locutions en cours de figement, mais il est relativement facile de revenir sur ce choix au besoin.

Lors de l'interrogation ou de l'analyse textométrique du corpus, la présence du balisage permet de retrouver assez facilement les occurrences concernées et d'ajuster éventuellement les calculs de fréquences[51]. À titre d'exemple, nous présentons ci-dessous quelques requêtes CQL exploitables sur un corpus balisé selon les exemples cités plus haut, après importation dans la plateforme TXM :

`[rend="aggl"]+ [rend!="aggl"]` : cette expression de requête permet de sélectionner toutes les séquences agglutinées (agglutinations simples ou locutions en cours de figement)

`[_.seg_type="fgmt"] expand to seg` : cette expression cible toutes les locutions en cours de figement.

50. http://goo.gl/EWZ7NC.
51. Bénédicte Pincemin, « Lexicométrie sur corpus étiquetés », dans Gérald Purnelle, Cédrick Fairon et Anne Dister (dir.), *Le Poids des mots*, *op. cit.*, t. II, p. 865-873.

Le cas précis de la segmentation en mots dans les éditions de textes, qu'elles se présentent sous forme papier ou électronique, a permis d'aborder de manière concrète quelques enjeux majeurs de l'essor du numérique dans le domaine de l'édition de textes et de ce qu'on peut désormais appeler la « philologie numérique ». Nous espérons avoir montré à partir de cet exemple particulier que le numérique offrait la possibilité d'inventer des manières de rendre compte, mieux que ne le faisaient les éditions traditionnelles ou de façon plus utile aux linguistes, de la réalité matérielle des manuscrits médiévaux. Ces technologies nouvelles offrent parallèlement des outils de recherche et d'analyse souples et puissants permettant de regrouper, de manier et d'analyser un volume de données très important tout en maîtrisant leur diversité.

La réflexion qui sous-tend les pratiques numériques émergentes vise à faire des choix de représentation du texte qui soient cohérents, consistants, documentés et aussi adaptés que possible aux objectifs scientifiques que l'on se donne. Ces choix incombent aux linguistes et aux philologues qui produisent et exploitent les ressources numériques. Ils permettent de renouer un dialogue actif entre spécialistes de disciplines récemment séparées. Ils sont également l'occasion de s'ouvrir aux sciences de l'information, et plus spécifiquement à l'informatique.

Les possibilités très riches offertes par le numérique ne doivent pas occulter le coût qui les accompagne. L'harmonisation et le partage des choix philologiques, linguistiques et techniques deviennent de ce fait une condition majeure des développements actuels et à venir. Les standards et les normes partagés, qui concernent les choix d'édition tout autant que les formats d'encodage, garantissent l'échange, l'enrichissement progressif, la diffusion et la pérennisation des ressources. Les licences libres qui sont utilisées de plus en plus largement, les licences *Creative Commons* en particulier[52], contribuent à la même démarche participative et communautaire. Ce mouvement

52. http://creativecommons.org.

très vaste rend la collaboration interdisciplinaire d'autant plus nécessaire pour la survie des disciplines elles-mêmes. Nous pensons qu'il est aussi et surtout un facteur de progrès et d'enrichissement réciproque.

Résumés

Julie GLIKMAN et Thomas VERJANS, Regards linguistiques sur les éditions de textes médiévaux

Résumé

Cette contribution constitue l'introduction du volume. Elle présente le contexte dans lequel ce numéro a été préparé et la volonté des directeurs du volume d'interroger les rapports entre les pratiques philologiques et les études de linguistique diachronique. Ces rapports peuvent se mesurer dans la place accordée aux faits linguistiques dans les introductions d'édition, ou inversement la place accordée aux variantes et à l'apparat critique dans les corpus numérisés. Elle présente ensuite les différentes contributions du volume.

Abstract

This contribution is the introduction to the volume. It presents the context in which this issue was prepared and the willingness of the editors to question the relationship between philological practices and studies of diachronic linguistics. These relationships can be evaluated by considering the importance given to linguistic facts in the introductory sections of editions. Conversely, it can also be evaluated by according to the importance given to variants and critical apparatus in digitized corpora. The various contributions of the volume are also introduced.

Nathalie BRAGANTINI-MAILLARD, Suivre la lettre du copiste : l'édition critique au service de la linguistique diachronique et diatopique. L'exemple du ms. Paris, BnF, fr. 99

Résumé

La connaissance des modalités d'évolution du français à la fin du Moyen Âge ne peut désormais s'affiner sans une reconnaissance véritable du rôle crucial que jouèrent les copistes au plan linguistique dans la diffusion et la survie des textes anciens. L'action du copiste est en effet double, en s'exerçant à la fois sur le plan horizontal de la circulation des textes d'un espace linguistique à un autre et sur le plan vertical de la transmission des textes à travers les époques. Dans la pratique scientifique, la prise en compte de cet apport déterminant doit passer non seulement par une édition des textes plus respectueuse de la version procurée par un manuscrit donné, mais aussi par un examen documenté, exhaustif et précis des phénomènes linguistiques qui particularisent les témoins retenus et les modifications de scribe. À terme, l'information rassemblée par ces profils linguistiques devrait permettre de mieux appréhender les phénomènes d'adaptation, de rajeunissement et d'enrichissement du français au Moyen Âge. À titre illustratif, nous nous proposons de montrer l'intérêt que présente le ms. BnF, fr. 99 pour suivre de manière privilégiée certains phénomènes de modernisation du français dans la seconde moitié du XV^e^ siècle, ainsi que l'influence que put exercer le lieu de copie occitanisant sur l'adaptation linguistique du texte, autrement dit les conditions d'échanges entre oïl et oc.

Abstract

Knowledge of how French evolved in the late Middle Ages can no longer be refined without a genuine recognition of the crucial linguistic role played by copyists in the dissemination and survival of ancient texts. Copyists act both on the horizontal dimension of the circulation of texts from one linguistic space to another, and on the vertical dimension of the transmission of texts through

the ages. This decisive contribution must be taken into account, not only by providing edition of the texts that are faithful to the version of a given manuscript, but also by a comprehensive and precise examination of the linguistic phenomena that characterize the witnesses and scribal modifications. Ultimately, these linguistic profiles will provide information for a better understanding of the phenomena of adaptation, rejuvenation and enrichment of French in the Middle Ages. To illustrate this, we examine ms. BnF, fr. 99, which displays exceptionally well certain phenomena of the modernization of French in the second half of the 15th century. It also demonstrates the influence that the place of copying with an affinity for Occitan may have had on the linguistic adaptation of the text, i.e. the conditions of exchange between Oïl and Oc.

Laurent BALON,
Pour une « troisième voie » en matière d'édition de textes d'ancien et de moyen français

Résumé

La pratique de l'édition de texte se trouve face à un dilemme : en partant des conseils trouvés dans les quelques articles sur la question et les manuels récents donnant des principes d'édition, on observe que les critères actuels de choix des variantes aboutissent à l'exclusion du matériau intéressant le linguiste qui, de son côté, aurait besoin d'un exposé intégral de toutes les données, sans tri. Ce besoin d'un non-choix est important, mais peu facile à satisfaire, voire impraticable à l'écrit, et la présentation des données intégrales du manuscrit se heurte à la lisibilité et à l'intelligibilité. L'objet de cette contribution est de présenter une méthode d'édition constituant un compromis entre l'édition critique traditionnelle et la transcription dite diplomatique, reposant sur un protocole de choix de variantes permettant de mieux satisfaire certains besoins des linguistes. Afin de fournir au linguiste des informations immédiatement exploitables et utiles à l'avancée de la discipline, le principe méthodologique proposé consiste à signaler dans l'édition

certains faits de langue relevant de la ponctuation du mot par l'emploi d'un code graphique qui en conserve la trace, à savoir un système de « tirets » déjà suggéré par Jacques Monfrin pour la transcription des documents d'archives, mais complété et appliqué pour la première fois à un texte littéraire par Nelly Andrieux-Reix. Le bien-fondé et l'intérêt de cette méthode seront illustrés par des études de cas en lien avec notre propre travail de recherche.

Abstract

Editors must cope with a dilemma: according to publishing principles in recent papers and textbooks, the current criteria for choosing variants excludes materials of great interest to linguists. They would need a comprehensive view of the data, without sorting. This is not easy to achieve, and even impossible on paper. The full presentation of the data of the manuscript hampers legibility and intelligibility. The purpose of this contribution is to present a compromise between traditional critical editing and diplomatic transcription, based on a protocol of choice of variants that better satisfies linguistic investigations. The proposed methodological principle aims at providing information that is immediately usable and useful for the advancement of the linguistics. This purpose is achieved by indicating facts relating to the punctuation of the word by using a graphic code that keeps track of them: a system of "dashes", suggested by Jacques Monfrin for the transcription of archival documents. This system is expanded and applied for the first time to a literary text by Nelly Andrieux-Reix. The merits and interest of this method will be illustrated by case studies related to our own research work.

Alexei LAVRENTIEV, Céline GUILLOT-BARBANCE et Serge HEIDEN, Enjeux philologiques, linguistiques et informatiques de la philologie numérique : l'exemple de la segmentation des mots

Résumé

Les linguistes travaillant sur l'histoire de la langue ont toujours exploité et utilisé comme principale source d'exploration les éditions « classiques », bien que depuis longtemps on connaisse leurs limites pour la recherche linguistique. Le développement des technologies modernes a d'un autre côté rendu le recours à de nouveaux outils (concordances, index, calculs statistiques) peu à peu indispensable à la recherche en langue, et plus récemment, les progrès continus de la technologie ont également permis d'envisager la réalisation d'éditions d'un nouveau type. L'édition numérique, qui a déjà donné lieu à plusieurs réalisations concrètes, a ainsi permis aux linguistes auparavant bridés par le papier et les techniques traditionnelles d'exprimer plus librement leurs besoins et leurs exigences. Plusieurs recherches récentes déjà publiées montrent l'efficacité de ce mouvement et le caractère novateur des acquis ainsi obtenus. À partir d'un exemple concret d'édition numérique interactive, notre présentation détaillera les enjeux méthodologiques liés à ces nouveaux outils et à ces nouvelles pratiques, en proposant une réflexion sur le concept de « philologie numérique » et en montrant ses principaux apports pour la recherche diachronique. Cette question sera illustrée en particulier par la question de la segmentation des mots.

Abstract

Linguists working on the history of language have always exploited "classical" editions as their main source of exploration, although the limits of such resources for linguistic research have long been known. On the other hand, modern technology has gradually offered new tools (concordances, indices, statistical calculations), that now prove to be indispensable. More recently,

the continuous progress has also made it possible to produce new types of editions. Digital publishing, which has already produced several achievements, has thus enabled linguists to express their needs and requirements better than before, freed from the constraints of paper and traditional techniques. Several recent studies demonstrate the efficiency of digital publishing and the innovative nature of the results obtained. Based on an example of interactive edition, we survey the methodological issues related to these new tools and practices, by investigating the concept of "digital philology", and by evaluating how it contributes to diachronic research. The specific issue of word segmentation will illustrate our point.

Nicolas MAZZIOTTA, L'activité éditoriale comme démarche de représentation de la connaissance : l'exemple de la ponctuation médiévale

Résumé

Cette contribution concerne le traitement éditorial de la ponctuation médiévale, selon une approche de la philologie comme activité de représentation des connaissances. Après une présentation des concepts de *connaissance* et d'*inscription* (des connaissances), le traitement de la ponctuation médiévale sert d'exemple aux questionnements que soulève toute activité éditoriale. Dans la démarche ecdotique, il s'agit d'identifier des classes de signes, pour distinguer ce qui est différent et rapprocher ce qui est similaire, mais également de segmenter correctement les unités observées. En outre, éditer consiste à « donner à lire », ce qui se manifeste par l'importance de choix ergonomiques importants pour garantir l'accessibilité de la connaissance inscrite. À bien des égards, l'inscription informatique de l'édition a beau ouvrir le champ des possibles, elle ne résout pas tout. Pour inscrire, il faut d'abord comprendre. L'édition ne pourra jamais se passer des *choix* foncièrement humains qui fondent le travail de construction de la connaissance.

Abstract

This contribution focuses on the editorial treatment of medieval punctuation, according to an approach of philology as an activity of *knowledge representation*. After a brief presentation of the concepts of *knowledge* and *inscription* (of knowledge), the treatment of medieval punctuation serves as an example for the questions raised by any editorial activity. Identifying classes of signs and distinguishing between what is different and what is similar are key parts of the ecdotic process. Moreover, by editing a text, one actually *makes it readable*. Consequently, ergonomic choices are prominent in this process, in order to guarantee the accessibility of the knowledge inscribed. In many respects, digital publishing opens up the field of possibilities, but it does not solve the fundamental problems. Understanding the text stands as the first step into building any valuable critical edition. Human *choices* will always remain the basis of any elaboration of knowledge.

Sylvie BAZIN-TACCHELLA et Gilles SOUVAY, Lemmatisation et construction automatique de ressources lexicographiques : les développements du lemmatiseur LGeRM

Résumé

Le lemmatiseur LGeRM, conçu à l'origine pour faciliter la consultation du *Dictionnaire du moyen français*, a connu depuis 2008 de nouveaux développements et est aujourd'hui utilisé dans de nombreux autres contextes, notamment dans l'interrogation de bases textuelles et la constitution de lexiques ou glossaires informatisés, autant d'outils qui peuvent servir d'aide à l'édition, le lemmatiseur ayant été intégré depuis à plusieurs grands projets d'édition en ligne. Cette contribution se propose de retracer l'histoire de la conception de LGeRM et de ses développements successifs, en montrant les différentes possibilités de l'outil illustrées à partir des projets récents.

Abstract

The LGeRM lemmatizer, originally designed to facilitate the consultation of the *Dictionnaire du moyen français*, has undergone new developments since 2008. It is now used in many other contexts. In particular, it helps the interrogation of textual bases and the constitution of computerized lexicons or glossaries. Additionally, the lemmatizer has also been integrated into several major online publishing projects in order to help the publishing process. This contribution retraces the history of the conception of LGeRM and its successive developments, by showing how recent projects make use of it.

Table des matières

Regards linguistiques sur les éditions de textes médiévaux
Julie Glikman & Thomas Verjans 7

Suivre la lettre du copiste : l'édition critique au service de la linguistique diachronique et diatopique. L'exemple du ms. Paris, BnF, fr. 99
Nathalie Bragantini-Maillard 17

Pour une « troisième voie » en matière d'édition de textes d'ancien et de moyen français
Laurent Balon 47

Enjeux philologiques, linguistiques et informatiques de la philologie numérique : l'exemple de la segmentation des mots
Alexei Lavrentiev, Céline Guillot-Barbance & Serge Heiden 77

L'activité éditoriale comme démarche de représentation de la connaissance : l'exemple de la ponctuation médiévale
Nicolas Mazziotta 103

Lemmatisation et construction automatique de ressources lexicographiques : les développements du lemmatiseur LGeRM
Sylvie Bazin-Tacchella & Gilles Souvay 121

Résumés/Abstracts 147