Les états anciens des langues à l'heure du numérique

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

ISBN : 979-10-231-0581-0

Maquette initiale : Compo-Méca
Réalisation : Emmanuel Marc Dubois (Issigeac)

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

pups@sorbonne-universite.fr
http://pups.paris-sorbonne.fr
Tél. (33) 01 53 10 57 60
Fax. (33) 01 53 10 57 66

Présentation

Joëlle Ducos
EA 4509 STIH
Université Paris-Sorbonne

Linguistique et informatique depuis plusieurs décennies font bon ménage : si les tableaux phonologiques ont correspondu à la binarité des premiers ordinateurs, les développements considérables du TAL, des lemmatisations, des annotations automatiques ont contribué à la fois à un renouvellement des outils d'analyse, avec une puissance considérable de traitement de la matière linguistique, mais aussi au développement de nouvelles approches, voire de nouvelles définitions ou de nouveaux concepts comme on peut le mesurer à la lecture des numéros 141 et 142 de l'*Information grammaticale* (2014) faisant le point sur la place de l'informatique et du numérique dans le traitement automatique de l'oral et de l'écrit. Ainsi Irène Tamba et Daniel Lugazzi tentent-ils de définir le *mot-numérique*, nouveau concept pour une réalité lexicale autre que le mot-forme ou le mot-lexème, un mot codé numériquement, qui ouvre des voies nouvelles à l'analyse linguistique et spécialement à celle du sens[1].

Nouvelles perspectives, révolution conceptuelle ? Il est certain que le numérique n'est pas du papier numérisé et consultable en ligne, en quelque sorte une pure évolution de support où le matériau papier disparaîtrait dans une image : les usages et les applications témoignent d'une transformation de fond, dans la méthode comme dans la réflexion. On pourrait penser qu'accoler l'adjectif *ancien* à *numérique* est paradoxal, voire antonymique. Or depuis que l'ordinateur est un outil pour

1. *Information grammaticale*, n° 142, 2014, p. 40-47.

les chercheurs, les nouvelles possibilités qu'il apportait ont tout de suite emporté leur adhésion, et spécialement celle des médiévistes, qu'ils soient historiens, linguistes ou littéraires. Citons la revue *Le médiéviste et l'ordinateur*, créée en 1979 par un groupe d'historiens et de chercheurs et publiée par l'IRHT, qui a contribué largement à la diffusion des perspectives liées à l'ordinateur, que ce soit dans la textométrie, comme appui à l'analyse des sources et des textes, ou par la création du site *Ménestrel*, à la fois plate-forme des ressources documentaires en ligne pour la médiévistique et lieu de publication pour les réflexions de chercheurs[2]. Les outils numériques ont été très vite utilisés pour les états anciens du français, ne serait-ce au début que par l'établissement de concordanciers, et par de multiples entreprises, soit de traitement de corpus, soit de dictionnaire. C'est ainsi que le corpus de Chrétien de Troyes, traité informatiquement par l'université d'Ottawa, a été le point d'origine du *Consortium international pour les corpus de français médiéval*, créé en 2004 par les universités d'Ottawa, du Pays de Galles, de Stuttgart, de Zürich, l'ENS-LSH, l'ATILF, et l'École nationale des chartes. Citons aussi la Base de français médiéval créée à l'initiative de Christiane Marchello-Nizia en 1989, et désormais projet phare du Laboratoire ICAR (UMR 5191 ENS LSH/CNRS). Pour les dictionnaires, c'est le *Dictionnaire de moyen français* (*DMF*) créé par Robert Martin qui apparaît comme la référence d'une entreprise conçue dès le départ sous un format informatique, transformant alors la lexicographie par l'idée d'un dictionnaire dont les versions successives soulignent les apports et les évolutions, en fonction d'un format qui n'est plus celui de la page éditée et qui est élaboré pour répondre aux contraintes et aux potentialités de la diffusion en ligne. Ces entreprises, qui reposent sur les textes en français médiéval, sont apparues alors que d'autres naissaient pour d'autres langues, dans d'autres pays, et reposaient sur les mêmes interrogations à propos de la place de l'informatique dans les disciplines de la

2. Ménestrel, en ligne : http://www.menestrel.fr/spip.php?rubrique397&lang=fr [consulté le 21 juin 2017].

médiévistique, qu'il s'agisse de la linguistique, de l'histoire, de la philologie ou la littérature. Ces travaux pionniers ont été à l'origine des réflexions actuelles et de la multiplication des projets numériques ou de linguistique outillée, que renforcent actuellement les financements nationaux et internationaux de la recherche. L'apport informatique y apparaît clairement comme une expansion considérable de potentialités de traitement, mais surtout comme un appel à la réflexion méthodologique et conceptuelle, ainsi qu'à la rigueur intellectuelle, qui fait progresser aussi dans la connaissance et dans l'analyse sémantique et lexicologique.

Il semblait donc nécessaire de procéder à une mise en perspective de l'apport du numérique à la connaissance des langues médiévales ou plus anciennes encore, qu'il s'agisse du latin, du français, de l'italien ou du grec, dans la mesure où le demi-siècle qui vient de s'écouler a permis la réalisation de plusieurs projets, anciens ou tout récents. Il s'agit de rendre compte d'expériences, de méthodes ou d'approches différentes, de tirer les leçons de ce qui relève désormais d'une histoire de la recherche informatique, de mesurer les résultats obtenus pour envisager l'étape suivante, à l'heure des *Big Data* et des *Linked Open Data* : représentation et modélisation des données qui permettent des interfaces et des requêtes multiples, gestion de ces données…

Le présent volume ne prétend pas répondre à tous les enjeux actuels issus aussi bien des nouvelles possibilités technologiques que de l'augmentation considérable des données disponibles. Il a pour ambition de mettre en évidence les convergences entre des projets séparés, sur des aires linguistiques différentes, menés par des chercheurs qui rendent compte de leurs méthodes, de leurs outils et de leurs avancées. Les applications sont multiples, et vont de la syntaxe à la stylistique en passant par les dictionnaires ou la sémantique. Loin de considérer en effet que le numérique est susceptible de détruire la diachronie et les études classiques, les expériences menées ces dernières années – et leurs résultats – prouvent combien il peut être le

moteur d'un renouvellement, à condition que la fin ne soit pas l'outil en lui-même, mais bien la recherche en linguistique : c'est ainsi que la nouvelle philologie a trouvé par l'édition numérique le moyen de mettre en œuvre de nouvelles formes d'édition et de rendre compte à grande échelle de la variance et de la variation. Qu'apporte le numérique ? Quelles en sont les limites, et quels nouveaux horizons ouvre-t-il ? Telles sont donc les questions sous-jacentes aux différentes contributions présentées ici. Elles témoignent toutes de l'importance des prémisses, de la nécessité de la mise au point d'une méthodologie rigoureuse, qui ne soit pas seulement technologique mais repose sur une connaissance véritable d'un état de langue. Elles révèlent aussi les difficultés, les nœuds éventuels qui peuvent se créer mais, surtout, elles témoignent de l'ouverture à de nouvelles dimensions, des ponts qui désormais sont praticables entre des domaines et des aires séparés alors que la technologie progresse, et permet d'aller bien au-delà de l'automaticité binaire des premières tentatives. Fouilles de textes, ontologies, lemmatiseurs, utilisation de standards internationaux, autant de perspectives qui se répondent et qui s'adaptent à la réalité mouvante et complexe des textes et des langues du Moyen Âge.

Ce numéro, issu du colloque final d'un programme ANR, CréaLScience (2010-2014)[3], qui a permis la conception du *Dictionnaire de français scientifique médiéval* (*DFSM*), se veut aussi une ouverture au-delà des frontières linguistiques et la mise en évidence d'une communauté de projets rendant compte de la diversité linguistique de la période médiévale. L'outil informatique permet ainsi de reconstruire la richesse médiévale, les réseaux, les relations et les échanges qui s'établissaient à cette époque par-delà l'obstacle de la langue. Cette communauté « numérique » qui se construit est assurément désormais le point de départ de nouvelles voies dans les territoires de la recherche médiévale.

3. En ligne : www.crealscience.fr.

À propos du *DMF* : réussites et pièges de la lexicographie électronique

Robert Martin
Académie des inscriptions et belles-lettres

La part du numérique ne cesse de croître – dans nos disciplines comme ailleurs. Le bénéfice est tel que l'on n'imagine pas un retour en arrière : un autre âge s'est ouvert, non pas épistémologique sans doute, les questions de fond restant inchangées, mais technique et méthodologique. Le stockage électronique des données, leur organisation, les accès instantanés que l'informatique autorise, les modèles interprétatifs et les représentations qu'elle suscite, la rigueur du contrôle qu'elle impose, tout cela est de si grande conséquence qu'il convient, avec le minimum de recul qui désormais s'instaure, d'en prendre une juste mesure. Si les avantages l'emportent, les pièges cependant ne sont pas inexistants. C'est vrai tout particulièrement en lexicographie. L'expérience du *Dictionnaire du moyen français* (*DMF*)[1] devrait en l'occurrence faciliter les entreprises similaires, qu'elles soient à leurs débuts ou déjà en cours. J'évoquerai tout d'abord ces avantages, avant d'insister sur les revers possibles, que l'expérience acquise devrait contribuer à déjouer.

Bénéfices de la lexicographie électronique

Dans l'histoire (encore récente) de la lexicographie électronique, trois grandes étapes dès à présent se dessinent : l'étape de l'informatique documentaire ; l'étape de l'informatisation des dictionnaires ; l'étape de l'élaboration

1. En ligne : www.atilf.fr/dmf.

lexicographique assistée par ordinateur. Chacune d'elles procure d'incontestables bénéfices.

La phase documentaire

Dans l'étape la plus ancienne, l'informatique se limite à la phase documentaire. Il s'agit alors :

– de rassembler, sous un format lemmatisé, un grand nombre de données ;
– de les affecter de références standardisées et immuables ;
– de les trier automatiquement selon des critères formels susceptibles d'en faciliter au mieux l'exploitation.

Fini le labeur fastidieux des fiches manuscrites, oubliées les références flottantes. Voyez le Godefroy : un même texte peut y être référencé sous des formes variables. Un exemple entre mille : *Le Racional des divins offices* de Guillaume Durand, adapté en français par Jean Golein, désormais accessible, du moins en partie, dans l'édition établie par Brucker et Demarolle. Ce texte figure dans Godefroy sous le nom d'auteur de « G. Durant » jusqu'à l'entrée *chapefol*, puis, à partir de l'entrée *collectaire*, sous l'étiquette bibliographique de « J. Goulain, *Ration.*, Richel. 437, p. ex., t. VII, 428b, *s.v. sincopiser* », ou « t. VII, 594a, *s.v. supererogation* » ; mais parfois aussi sous la forme « J. Goulain, *Trad. du Ration. de G. Durant*, B.N. 437, p. ex. GDC X, 652a, *s.v. segregation* ». Des références bibliographiques immuables et, grâce à l'informatique, commodes d'accès, évitent ces flottements.

Les opérations de tri, même élémentaires, aident à dominer la masse des informations. Ainsi, dans le *Trésor de la langue française* (*TLF*), dès 1966 (il y a près d'un demi-siècle !), le programme dit des « groupes binaires » a permis de donner une idée plus juste de la syntagmatique ; la démarche est simple : *maison* apparaît dans le corpus avec une certaine fréquence ; de même *campagne* ; on compare alors la probabilité de trouver *maison* et *campagne* côte à côte (comme dans *maison de campagne*) par le seul fait du hasard à la fréquence effective, en l'occurrence significativement supérieure ; le seul hasard ne

pouvant expliquer la fréquence de *maison de campagne*, il ne peut s'agir que d'un fait linguistiquement pertinent ; c'est là une donnée qui ne peut laisser le lexicographe indifférent. Les tris de cette espèce remontent aux débuts de l'informatique : ils ont marqué la discipline.

L'informatisation

La seconde phase de l'histoire de la lexicographie électronique est celle de l'informatisation des dictionnaires, d'abord par rétroconversion (comme pour le *Dictionnaire d'Oxford* ou pour le *TLF*), ensuite par balisage initial (comme pour le *DMF*). Les avantages en sont désormais si connus et si unanimement appréciés qu'il suffit de les rappeler brièvement sous quelques rubriques ; ils tiennent à la diversité des *accès* et à la commodité des *liens*.

Les *accès* sont en effet très divers :

– ils sont déliés de la linéarité (on peut afficher toutes les occurrences de *campagne*, même en dehors de l'article *campagne*) ;
– ils peuvent engager une lemmatisation (dans le *DMF*, l'accès dit « Mot ou forme » propose l'ensemble des articles auxquels une forme peut théoriquement appartenir : ainsi en demandant *flageole*, on obtient le lemme *flageole*, substantif qui existe, mais aussi *flageoler*, dont *flageole* est une forme fléchie ; il va sans dire que la pertinence des propositions faites par le lemmatiseur ne saurait être absolue ; cependant les propositions correctes sont à présent nettement supérieures à 95 %, et proches de 99 % si l'on tient compte des réponses plurielles parmi lesquelles figure la réponse correcte) ;
– ils peuvent s'opérer par fragments (dans le *DMF*, sous l'intitulé « Filtre » : on peut afficher par fragments initiaux, p. ex. les mots qui commencent par *in-* ; par fragments terminaux, p. ex. tous les mots qui se terminent par *-tendre* : *attendre*, *contendre, contrattendre, détendre, distendre, entendre, forestendre, mesentendre, parentendre, partendre, pourtendre, prétendre, protendre, retendre*… ; ou par

fragments internes, p. ex. tous les mots qui contiennent *-fil-* : *affiler*, *défiler, effiler*, *effiloir, enfiler, forfiler, profiler, pourfiler*, *refiler*…) ;

– les accès peuvent aussi se réaliser par types d'informations, p. ex. par l'étymon, par la syntagmatique ou par le repérage sous une balise donnée, p. ex. *maison* ou *campagne* dans les définitions.

Dans le *DMF*, les *liens hypertextuels* se diversifient en trois types de liens :

– des *liens internes*, comme le lien des références abrégées avec la Bibliographie, ou bien le lien d'une forme quelconque dans une citation avec le ou les articles qui en traitent ;
– des *liens avec les Bases* qui ont servi à construire le Dictionnaire, pour le *DMF* avec la Base des « Lexiques préalables » et avec les « Bases textuelles » ;
– des *liens avec d'autres ouvrages* ; ainsi le *DMF* permet, article par article, d'ouvrir le *TLF*, le Godefroy ou le *FEW* (en 2015 l'*AND*, le *DECT* et en partie le *DEAF*).

L'élaboration par voie électronique

La troisième phase est celle de l'élaboration du Dictionnaire par voie électronique, celle d'une lexicographie assistée par ordinateur. Grâce à Gilles Souvay, le *DMF* dispose de divers outils qui assistent le rédacteur tout au long de sa démarche.

Ainsi le *DMF* s'élabore selon une *grammaire lexicographique* qui garantit l'homogénéité de l'écriture. Cette grammaire est un système qui, au fil de la rédaction, spécifie le type d'information qu'il convient de fournir ; elle affiche au fur et à mesure les balises à remplir et les choix à faire parmi les balises possibles à l'endroit où l'on est arrivé. Ainsi la première balise, obligatoire, est celle de la vedette : le curseur indique où il convient de l'écrire ; la forme s'enregistre automatiquement en gras ; vient ensuite un premier choix : on peut se borner à un simple renvoi (qui doit être précisé à l'endroit où le curseur s'est placé) ou bien on choisit de poursuivre, auquel cas, un exposant est possible (en cas d'homonymie), puis c'est le « code grammatical » qu'il