Journées Meditext – 18-19 mai 2011

Compte rendu

Ces journées avaient pour objectif de présenter les premiers développements d'un logiciel destiné au traitement des textes médiévaux, conçu dans le cadre du programme européen *Signs and States* dirigé par Jean-Philippe Genet, de les confronter aux autres projets actuellement menés et de nourrir la discussion autour des problèmes soulevés par ce traitement.

MERCREDI 18 MAI

Après avoir présenté l'ensemble du projet *Signs and State*[1], Jean-Philippe Genet a défini plus précisément le programme Meditext qui comprend deux volets : le premier est celui de la constitution d'un corpus de textes médiévaux, essentiellement centré sur les textes politiques et résolument multilingue. Pour l'instant, le corpus comprend des textes en anglais, français et latin, mais une extension est envisagée vers le castillan et l'italien. La forme de ces textes présente nombre de difficultés : l'orthographe n'est pas fixée et les langues sont extrêmement fluides. De fait, la question de la lemmatisation se pose de façon très différente pour la période médiévale et pour la période contemporaine. Mais les historiens ne peuvent traiter ces textes comme les philologues. L'objectif du second volet du programme est donc de mettre sur pied un outil permettant de constituer un corpus dans de bonnes conditions et de les traiter avec un logiciel de textométrie. Il ne s'agit donc pas d'un projet d'édition électronique, mais de la mise à disposition de textes au format brut.

Chris Fletcher, Rachel Moss (pour l'anglais), Naomi Kanaoka (pour le français) et Lauren Henras (pour le latin) présentent ensuite le corpus tel qu'il est actuellement constitué[2]. La base du corpus actuel est un corpus de textes d'origine anglaise et française rassemblés depuis les années 1970 sous la direction de Jean-Philippe Genet et Claude Gauvard. Il comprend essentiellement des textes « politiques », qu'ils aient trait à des événements politiques identifiés ou qu'ils soient plus généralement consacrés au bon et au mauvais gouvernement ; des textes gouvernementaux (des proclamations et ordonnances par exemple) et des textes adressés au roi par ses sujets (cahiers de doléances ; requêtes ; lettres de rémission). Le corpus en anglais, dont les premiers textes datent de la fin du XIV^e^ siècle, comprend actuellement 650 000 mots. Les corpus français et latins ne sont pas encore suffisamment structurés quantitativement par rapport au corpus anglais (le corpus français, par exemple, compte environ 180 textes ou groupes textuels). Un gros travail

---

[1] Un résumé du programme est disponible sur le site du LAMOP : http://lamop.univ-paris1.fr/spip.php?rubrique149.

[2] Voir la liste des textes du corpus en fichier joint.

pour normaliser et compléter les textes manquants est en cours. Insistance est faite sur les problèmes d'harmonisation des textes, dont les origines sont très diverses, sur la variabilité des normes informatiques, ainsi qu'à la difficulté de travailler sur des éditions anciennes.

Mourad Aouini s'attache ensuite à l'interface informatique en cours de constitution, pour l'instant intitulée PALM (Plateforme d'Analyse Linguistique Médiévale)[3] et explique les présupposés de la méthode hybride pour l'étiquetage morphosyntaxique qui a été adoptée. L'étiquetage morphosyntaxique est une opération qui permet, à partir d'un ensemble de couples (forme, étiquette morphosyntaxique), de choisir, pour chacun des mots du texte parmi ses étiquettes morphosyntaxiques associées celle(s) qui correspond(ent) au contexte[4]. Les méthodes d'étiquetage morphosyntaxique se divisent en deux grandes approches :

- une approche à base de règles qui utilise généralement des « grammaires locales »[5] pouvant être modélisées sous forme d'automate ou de transducteur ;
- une approche probabiliste, la plus utilisée actuellement, qui peut être mise en œuvre à l'aide de modèles de Markov cachés ou à l'aide d'arbres de décision, TreeTagger par exemple.

Ces approches classiques suffisent-elles pour étiqueter des textes médiévaux ? Une étude réalisée par E. Tzourkermann en 1996 sur deux corpus en français moderne extraits du journal *Le Monde* a montré que plus de la moitié des mots ne sont pas ambigus et qu'une grosse part de l'ambiguïté est détenue par un petit nombre des mots fréquents, généralement des mots outils[6]. Mais si le taux d'ambiguïté se limite à un nombre fini de mots pour les langues modernes, il existe, pour les langues médiévales, un grand nombre de possibilités pour une même forme morphosyntaxique et une liste complète des formes est impossible à établir[7].

Afin de répondre aux spécificités et aux diversités des langues médiévales, il est donc proposé une méthode hybride qui combine un tagger à base des règles réalisées en utilisant les normes, modules, automate et transducteur du système Intex, et la puissance des étiqueteurs probabilistes telles que TreeTagger et Mate. Cette méthode est indépendante de la langue et peut être appliquée à plusieurs langues si l'on dispose des ressources linguistiques nécessaires (corpus étiqueté, dictionnaires, règles de morphologie, etc.).

Les discussions qui ont suivi ces présentations ont été nourries et plusieurs problèmes ont été soulevés, en particulier :

---

[3] Voir le document de présentation en pièce jointe.

[4] Serge Fleury, site plurital : http://www.tal.univ-paris3.fr/cours/masterproj.htm

[5] Maxi Silberztein, manuel d'Intex : http://mshe.univ-fcomte.fr/intex/downloads/Manuel.pdf

[6] Benoît Habert, Adeline Nazarenko et André Salem, *Les linguistiques de corpus*, Paris, 1997.

[7] Gilles Souvay et Jean-Marie Pierrel, « Lemmatisation des mots en moyen français », *Traitement automatique des langues*, 50, 2009, en ligne à l'adresse suivante : http://www.atala.org/LGeRM

- la question de la possibilité d'une interrogation multilingue – sachant, qu'il existe de nombreux termes communs, en particulier à l'anglais et au français (David Trotter, université d'Aberystwith) ; pour l'instant, le travail est accompli langue par langue, mais avec l'idée de pouvoir réaliser de véritables comparaisons, notamment dans le cas de textes présents en plusieurs langues (comme le *De Regimine principum* de Gilles de Rome).
- le problème d'une erreur d'identification originelle du terme (Bertrand Guelf, ENC) ; il ne se pose que dans une moindre mesure, car les textes font l'objet d'un nettoyage attentif et que la plateforme permet aussi de déceler des erreurs. Il est cependant probable qu'à terme, la constitution d'un dictionnaire de règles améliorant le repérage de ces erreurs devra être envisagé.
- la question sensible de la segmentation des mots (Serge Heyden, ENS Lyon) ; pour l'instant, la plateforme ne permet pas de resegmenter un texte après coup, mais une réflexion doit être menée sur cette question.
- la question de la distinction entre bibliothèque et corpus (Chris Fletcher) ; l'objectif est que chaque historien puisse se constituer son propre corpus au sein de la bibliothèque (ou en ajoutant des textes à cette dernières). À cet égard, différents niveaux de droits d'utilisateurs ont été prévus.
- la question du dictionnaire du lemmatiseur ; Francesco Stella (Sienne) note qu'il existe à Francfort un projet assez proche, dirigé par Bernhard Jussen, sur les textes allemands et latins, qui prévoit un troisième niveau de lemmatisation pour constituer un dictionnaire de signifiés dans les différentes langues.
- pour le moyen anglais, la question est particulièrement délicate, car les graphes Intex ne fonctionnent pas, la différence avec l'anglais moderne étant trop importante. Il est peut-être possible, cependant, de récupérer les données du *Linguistic Atlas of Late Medieval English*[8].

Un second temps de l'après-midi a été consacré à la présentation d'autres projets de traitements linguistiques. Nicolas Mazziota (Liége) a d'abord présenté le projet *Syntactic Reference Corpus of Medieval French* (SRCMF) dirigé par Achim Stein et Sophie Prévost, et dont l'objectif est l'analyse syntaxique de textes en ancien français (France, IX^{e}-XIV^{e} siècles) afin de comprendre le processus de construction des phrases[9]. Il porte sur deux corpus d'environ 3 millions de mots chacun issus de la Base de Français médiéval[10] et du Nouveau Corpus d'Amsterdam[11]. L'annotation manuelle de ces textes est assurée par le biais d'un logiciel développé par Nicolas

[8] A. MacIntosh, M. L. Samuels, M. Benkin, assistés de M. Laing et K. Williamson, *A Linguistic Atlas of Late Medieval English*, 4 vol., Aberdeen, 1986.
[9] Le projet peut être consulté à l'adresse suivante : http://www.lattice.cnrs.fr/Projet-ANR-Syntactic-Reference.
[10] http://bfm.ens-lyon.fr/.
[11] http://www.uni-stuttgart.de/lingrom/stein/corpus#nca.

Mazziota et intitulé Nota Bene, fondé sur un modèle syntaxique dépendanciel[12]. Il permet des annotations multiples et des vues plurielles pour l'annotateur, l'objectif ayant été de construire un logiciel ergonomique pouvant être rapidement pris en main. Les procédures d'annotations sont cependant drastiques dans la mesure où ces dernières sont manuelles. Elles se déroulent en trois étapes : 1. Annotation parallèle aveugle ; la démarche consiste à identifier la phrase, puis le nœud verbal qui fonde la phrase, et enfin les dépendants du nœud verbal qui sont caractérisés et délimités. C'est une démarche récursive. 2. Correction croisée aveugle ; le logiciel permet de comparer les analyses d'annotateurs différents et de faire ressortir les divergences. 3. Contrôle et décision finale. Pour l'heure, environ 500 000 mots ont été annotés, dont la moitié ont été corrigés. Différents formats de sortie et d'exploitation sont possibles (tel quel, Graphviz, TigerXML, CoNLL). À terme, il est envisagé d'intégrer Nota Bene dans la plateforme TXM.

La discussion sur l'intervention de Nicolas Mazziota est ouverte par Olivier Bertrand (ATILF, Nancy) sur le problème de la délimitation des phrases – aussi sensible que celui de la segmentation des mots… En ce qui concerne le projet SRCMF, la segmentation des éditeurs a été respectée. Mais dans tous les cas, ce sont des questions à soulever dès l'origine pour toute entreprise de ce type.

Nicolas Mazziota revient ensuite sur la question des possibles usages de ces analyses syntaxiques par les historiens. Pour les chartes par exemple, il est ainsi possible de rechercher les acteurs, les clauses spécifiques, mais aussi de s'attacher à l'évolution phraséologique.

Après cette discussion, Serge Heiden présente la plateforme TXM (logiciel open-source développé sous Windows, Linux et Mac et fonctionnant également en ligne)[13], qui est davantage orientée vers la morphosyntaxe que vers la lemmatisation. Contrairement au projet Meditext, une grande attention est portée à l'aspect éditorial. TXM traite en priorité des corpus XML-TEI (ce qui est le cas de la Base de Français médiévale, encodée en XML-TEI P5), mais aussi des corpus XML et des corpus TXT. Les analyses développées dans TXM sont à la fois qualitatives (avec un moteur de recherche profond et l'utilisation de KWIC Concordance[14]) et quantitatives (analyses

[12] Voir N. Mazziota, « Concepts et modèle de données du logiciel NotaBene. Spécification technique », 2010, en ligne à l'adresse suivante : http://orbi.ulg.ac.be/handle/2268/11392 ; idem, « Building the Syntactic Reference Corpus of Medieval French Using NotaBene RDF Annotation Tool Langue du document », 2010, en ligne à l'adresse suivante : http://orbi.ulg.ac.be/handle/2268/34505.

[13] http://textometrie.ens-lyon.fr/.

[14] http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic_e.html.

factorielles, spécificités, cooccurrences…). La plateforme TXM a été développée dans le cadre du projet Textométrie mais s'intègre désormais dans de nouveaux contrats (en particulier l'Equipex Matrice). Un format spécifique XML-TXM est en cours d'élaboration, reposant sur un modèle de données liant des unités documentaires et leurs métadonnées (élément <text>) et des unités lexicales et leurs propriétés (élément <w>). Une réflexion est également menée pour mettre en place une stratégie fine de « tokenisation », permettant de résoudre les problèmes soulevés par la gestion des textes bruts lorsque ces derniers subissent des transformations (cela peut être le cas pour les transcriptions de performances orales ou lorsque des changements de paradigmes interviennent dans les éditions critiques).

Les discussions subséquentes portent essentiellement sur la question des modèles linguistiques (jeux d'étiquettes), qui reposent sur les corpus d'apprentissage.

La journée du JEUDI 19 MAI s'est déroulée sous forme de différentes tables rondes.

TABLE RONDE N°1 – Les logiciels et la lemmatisation (modérateur : Stéphane Lamassé, LAMOP)

• Gilles Souvay (ATILF, Nancy) a tout d'abord présenté son logiciel de lemmatisation, intitulé LGeRM[15], développé dans le cadre du *Dictionnaire de Moyen Français*[16] et d'autres programmes de recherches. Il s'agit d'un analyseur hors-contexte, fondé sur un dictionnaire de règles (4500 au total, dont 4000 concernant l'inflexion verbale). Si le traitement des ambiguïtés est problématique (30% de mots ambigus en moyenne), il est précieux pour la richesse de la base de connaissances et pour son adaptation à différents états de langue. Un projet interne de l'ATILF vise à sa réécriture et à une diffusion multiplateforme.

• Frédéric Glorieux (ENC, Paris) présente ensuite les expériences actuellement réalisées sur l'édition électronique du *Glossarium mediæ et infimæ latinitatis*, autrement dit le Du Cange[17]. L'usage d'un outil de lemmatisation, du même type que le LGeRM, vise à pouvoir effectuer des recherches au sein de cette édition, les caractères jokers étant trop limités. Il est fondé sur le protocole hunspell, employé pour les correcteurs orthographiques d'Openoffice. Un corpus d'expérimentation de 50000 occurrences et de 13000 formes a été constitué. L'équivocité reste importante (environ 20%), mais les ressources sont enrichies.

Une discussion s'engage alors sur les questions de la qualité et de la capitalisation des ressources linguistiques, et plus généralement du partage des ressources. Cela pose des problèmes de standardisation importants mais aussi des questions de licence et de droit, qui ne sont pas encore vraiment résolues.

---

[15] Pour une présentation détaillée, voir G. Souvay et P. Pierrel, « LGeRM. Lemmatisation des mots en moyen français », TAL, 50, 2009, p. 149-172, disponible à l'adresse suivante : http://www.atala.org/LGeRM.

[16] http://www.atilf.fr/dmf/.

[17] http://ducange.enc.sorbonne.fr/.

• Kim Gerdes (Sorbonne Nouvelle) revient ensuite sur l'analyse dépendantielle et les arbres de dépendance ordonnés qui peuvent être projetés linéairement sur une phrase. Il présente son logiciel Vakyartha[18], un annotateur en ligne utilisable pour n'importe quelle langue, utilisé notamment pour le projet Rhapsodie sur le français parlé[19].

• André Salem (Sorbonne nouvelle), concepteur du logiciel Lexico 3[20] soulève ensuite la question plus générale des unités en analyse statistique, au sein des deux options principales actuellement dessinées, la philologie numérique et la textométrie. Il observe tout d'abord qu'il faut encore renforcer les réflexions sur la ligne rouge entre les données numériques et ce qu'on y voit ; puis revient sur la question des unités de segmentation – qui n'ont pas toujours la même valeur selon les langues (par exemple en russe, en allemand ou en arabe). Il avance la notion de « types généralisés », c'est-à-dire des occurrences de ce qui n'est pas forcément défini par une forme graphique ou par un dictionnaire, par exemple tous les mots commençant par « négo ». Leur emploi est particulièrement intéressant dans le cadre de la topographie textuelle, des calculs de distance et de projections. Au final, l'idée est d'arriver à un certain consensus sur la question des unités (mot, *token*…).

• William Martinez (Sorbonne Nouvelle) présente ensuite un cas concret de traitement d'un corpus contemporain, le corpus Silicose (1800-1980), constitué dans le cadre du projet de recherche « Étude transnationale d'une maladie professionnelle exemplaire : la silicose et la santé au travail en France et dans les pays industrialisés » (ANR SEST)[21] ; il s'avère que dans le cas présent, la lemmatisation est plutôt contreproductive.

A suivi une discussion sur les vertus et les vices de la lemmatisation, les intervenants soulignant cependant que les enjeux diffèrent profondément selon qu'il s'agit d'une langue contemporaine ou d'une langue ancienne.

Les trois tables rondes suivantes ont eu pour objet de dresser un état des lieux et de dessiner les perspectives des programmes actuels concernant les différentes langues présentes dans Meditext.

TABLE RONDE N°2 – Le point sur la linguistique de corpus et le latin (modérateur : Monique Goullet, LAMOP)

---

[18] http://arborator.ilpga.fr/vakyartha/.

[19] http://rhapsodie.risc.cnrs.fr/fr/.

[20] http://www.tal.univ-paris3.fr/lexico/.

[21] Voir le dossier « La silicose, un cas exemplaire », *Revue d'histoire moderne et contemporaine*, 56, 2009/1, p. 83-176.

• Dominique Longrée, directeur du Laboratoire d'analyse statistique des langues anciennes (LASLA, Liège), présente tout d'abord les réalisations et les projets d'un des plus anciens laboratoire d'analyse des langues anciennes (il a été fondé en 1961)[22]. Tourné à l'origine vers la constitution de bases textuelles, principalement littéraires, il s'oriente actuellement vers des outils relevant du traitement automatique des langues. Actuellement, la base comprend deux millions de mots de latin classiques, ainsi que 350 000 mots de médio-latin et de néo-latin. Les textes sont complètement lemmatisés, pourvus d'une analyse morphologique complète, vérifiée par des philologues. Le codage est de type alphanumérique, et il est possible de travailler à la fois sur des formes graphiques, des formes standardisées et des lemmes. Le lemmatiseur s'appuie sur un dictionnaire et des règles. Actuellement, les fichiers sont en cours de migration vers des chantiers XML, en vue d'une exploitation sur la plateforme TXM. À cette fin, il est prévu la création d'une interface d'encodage (en collaboration avec l'ATILF), et du couplage du lemmatiseur/analyseur du LASLA avec le Tagger TnT[23]. L'insertion d'analyses syntaxiques est également en projet, qui peut s'avérer précieux pour l'historien.

• Caroline Philippart (Liège) présente ensuite un exemple de lemmatisation d'un corpus de textes hagiographiques du Haut Moyen Âge, qui présentent une grande diversité dans la langue et pour lesquels il est impossible d'effectuer la correction des variantes sur les bases de normes classiques. Il a donc été nécessaire d'adapter les protocoles du LASLA à ce corpus, ce qui est relativement aisé grâce au codage alphanumérique, bien que différents problèmes se soient posés, en particulier pour les changements de déclinaisons, de conjugaison ou la syntaxe.

• Olivier Canteaut (ENC, Paris) est revenu sur les enjeux de l'édition électronique et sur les usages des applications liées à cette dernière. L'École nationale des Chartes met en ligne à la fois des éditions originales et des numérisations d'éditions du XIX^{e} siècle. L'objectif de ces dernières est surtout de mettre à disposition des corpus préparatoires aux ou complémentaires des éditions originales. C'est le cas, par exemple, pour les cartulaires de l'Île de France[24], qui permettent d'offrir des points de comparaison avec le cartulaire blanc de Saint-Denis[25]. Afin de pouvoir effectuer des recherches au sein de ces éditions, elles ont été océrisées et un dictionnaire de formes adapté à l'état de langue du document traité a été constitué. Mais l'ENC aborde depuis quelques temps la question d'une recherche lemmatisée, y compris pour les noms propres, sachant que les outils actuels sont assez peu performants pour les textes médiévaux. Il est d'abord essentiel de constituer des ressources. Le projet Omnia, en partenariat avec l'ARTeHIS (Dijon) et l'IRHT, a

---

[22] http://www.cipl.ulg.ac.be/Lasla/.
[23] http://www.coli.uni-saarland.de/~thorsten/tnt/.
[24] http://elec.enc.sorbonne.fr/cartulaires/
[25] http://elec.enc.sorbonne.fr/cartulaireblanc/

pour objet, à partir du Du Cange, la constitution d'une base de lemmes et de variantes, automatiquement fléchies pour produire un dictionnaire de formes[26].

Le choix a été fait de proposer un outil qui offre un taux de rappel maximal, avec un jeu d'étiquette minimal (une dizaine) et une base de lemmes aussi réduite que possible (regroupement des dérivés, préfixes, suffixes…). La conséquence en est un « bruit » relativement élevé ainsi qu'un taux d'ambiguité important, mais le taux de reconnaissance est actuellement de 80% (phase d'entraînement).

• Benoit Grévin (LAMOP) s'est ensuite interrogé sur la possibilité d'arriver à une approche informatique du latin politique solennel à la fin du Moyen Âge, caractérisé par une formularisation extrême. On retrouve en effet des mécanismes semi-automatiques (_x_ _ _ _x_ par exemple, où x = accent) dans des dizaines de milliers de lettres. La synonymie apparaît comme un mécanisme particulièrement important. Mais d'autres sont bien plus subtils, en particulier dans le domaine des variations qui, par rapport à deux pôles de sémantisme, peuvent être pratiquement infinies. Autre problème majeur, il s'agit d'un corpus ouvert encore à créer, mais qui doit être le plus large possible : on compte peut-être entre 500 000 et 1 million de textes écrits sous influences… c'est-à-dire à peu près l'ensemble des diverses productions des notaires de l'Europe occidentale à la fin du Moyen Âge. Au final, Benoit Grévin s'interroge sur le fait qu'il s'agirait d'une analyse à mi-chemin entre linguistique et stylistique.

Il apparaît dans la discussion que certains, en particulier au LASLA, se sont justement engagés dans une réflexion sur la modélisation des motifs.

TABLE RONDE N°3 – Le point sur la linguistique de corpus et l'anglais (modérateur : Aude Mairey, LAMOP)

• Mark Ormrod (York) rappelle tout d'abord qu'il faut compter en ce domaine avec une forte tradition anglaise de résumé et de classement ; surtout, une grande partie des projets britanniques portent sur des textes qui ne sont pas en anglais. À partir de son expérience de directeur du programme *Medieval Petitions*[27], Mark Ormrod se demande si cette masse de documents numérisés peut conduire l'historien à de nouveaux questionnements. Il souligne trois aspects principaux :

- un aspect paléographique, étudié par Lynn Mooney, concerne en premier lieu l'identification de certains des scribes, très importante pour mieux cerner les contextes de production des manuscrits mais aussi pour étudier les dynamiques politiques des pétitions. Les premiers résultats

[26] Voir B. Bon, « OMNIA – Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins », BUCEMA, 13, 2009, p. 291-292, en ligne à l'adresse suivante : http://cem.revues.org/index11086.html.

[27] http://www.nationalarchives.gov.uk/documentsonline/petitions.asp.

ont conduit à l'identification d'un certain nombre de scribes londoniens qui travaillent aussi pour la municipalité ou pour des guildes importantes, et qui copient des textes littéraires.

- un deuxième aspect, mené par Gwilym Dodd, concerne la forme des pétitions et les conventions diplomatiques.

- un troisième aspect enfin, dont s'occupe plus précisément Mark Ormrod, envisage la rhétorique des pétitions et les usages de langages politiques particuliers (ceux de la justice et de la grâce, et particulièrement celui du bien commun). Se pose le problème des origines et des modèles d'acquisition du langage politique.

• Keith Williamson (Edinburgh) présente ensuite le *Linguistic Atlas of Early Middle English*[28] et le *Linguistic Atlas of Older Scots*[29], héritiers du *Linguistic Atlas of Late Middle English*, fondé sur un système de questionnaires linguistiques. Pour les versions en ligne du LAEME et du LAOS, un système de corpus taggé a finalement été préféré, d'autant que les ressources textuelles étaient bien moins importantes, surtout pour le LAEME.

• Odile Piton (Paris 1) évoque pour sa part, ses expériences de lemmatisation, par le biais du logiciel Nooj[30], concernant l'anglais du XVII[e] siècle, qui ont conduit à la constitution d'un dictionnaire des formes de cet état de langue et des ressources linguistique tels que des règles morphologique, syntaxique et contextuelles sous formes des automates et des transducteurs[31].

• Rachel Moss (LAMOP), rappelle enfin avec force quelques enjeux soulevés par l'étude du moyen anglais : l'importance de la culture multilingue dans laquelle il s'inscrit, surtout si l'on se positionne – et c'est indispensable – par rapport aux utilisateurs, et non par rapport à une langue abstraite ; et les problèmes liés à la division traditionnelle, dans le monde anglo-saxon, entre les études linguistiques et les études littéraires – division que le programme Meditext cherche en partie à combler.

TABLE RONDE N°4 – Le point sur la linguistique de corpus et français (Modératrice : Claude Gauvard, professeur émérite)

Claude Gauvard effectue tout d'abord un rappel sur la préhistoire de l'informatique et les origines de Meditext, insistant sur l'intérêt qu'il présente pour comparer des corpus dans le domaine du langage politique.

---

[28] http://www.lel.ed.ac.uk/ihd/laeme1/laeme1.html.

[29] http://www.lel.ed.ac.uk/ihd/laos1/laos1.html.

[30] http://www.nooj4nlp.net/pages/nooj.html.

[31] Voir H. Pignot et O. Piton, « Étude et traitement automatique de l'anglais du xviie siècle : outils morphosyntaxiques et dictionnaires », JLC, 6, 2009, en ligne à l'adresse suivante : http://web.univ-ubs.fr/corpus/jlc6.html.

• David Trotter (Aberystwyth) présente les différents corpus existant pour l'anglo-normand, outre ceux déjà mentionnés par Mark Ormrod (rouleaux parlementaires, pétitions) : le corpus en ligne de correspondances anglo-normandes compilé par Richard Ingham et Ruth Vorstman[32] ainsi que l'*Anglo-Norman Year Books Corpus*, regroupant les *Years Books* (recueils juridiques) de la fin du XIII^e^ siècle à la fin du XIV^e^ siècle[33]. Il souligne que ces corpus soulèvent un problème de corrélation entre oral et écrit – les lettres, comme les *Year Books*, sont censées être des transcriptions, mais sont en fait hautement formalisées. David Trotter évoque ensuite la base de l'*Anglo-Norman Dictionary*[34], qui rélève un certain nombre de problèmes cruciaux, en particulier sur la question des droits d'auteur : nombre de textes n'ont pu être inclus dans la base de l'AND étant sous copyright, ce qui constitue un véritable problème scientifique.

• Alexei Lavrentiev, (ICAR, Lyon) s'attache davantage aux corpus de français continental. Après avoir brièvement fait l'historique des premiers corpus (notamment Frantext, le Corpus d'Amsterdam, le DMF), il insiste sur l'importance du Consortium pour les corpus de français médiéval (CCFM)[35], qui regroupe nombre d'institutions intéressées par la question (Université d'Ottawa, École Normale Supérieure Lettres et Sciences humaines, Université de Stuttgart, Université de Zürich, Laboratoire ATILF, Université du Pays de Galles, École nationale des chartes). Il présente ensuite de manière plus détaillée le Nouveau Corpus d'Amsterdam, conçu en 1987 par Anthonij Dees, qui rassemble plus de 300 échantillons de textes et plus de 3 millions de mots taggés, lemmatisés et téléchargeables ; la base Textes de français ancien du laboratoire d'Ottawa ; la Base de Français médiéval… Là encore, se pose le problème de l'accessibilité et des droits d'auteurs. En ce qui concerne la BFM, par exemple, seul un quart des textes est accessible.

• Olivier Bertrand, enfin, présente son projet européen en cours, *Histoire du lexique politique français*, qui comprend deux volets principaux, d'une part l'édition critique de la première traduction de la Cité de Dieu en français par Raoul de Presles (1371-1375) et d'autre part des études diachroniques sur la question de lexicalisation, ce qui nécessite un corpus très large.

La discussion qui suit revient, pour terminer, sur les questions de fond des droits d'auteur et des relations avec les éditeurs commerciaux qui sont revenues dans la plupart des présentations des différents corpus.

Aude MAIREY

---

[32] http://wse1.webcorp.org.uk/anglo-norman/.

[33] Ce corpus n'est pas en ligne, mais il est possible d'écrire à Pierre Larrivée (p.larrivee@aston.ac.uk) pour y accéder.

[34] http://www.anglo-norman.net/.

[35] http://ccfm.ens-lyon.fr/.