Quelles perspectives pour la textométrie des états des langues passées ?

Aude Mairey

Aude Mairey

Quelles perspectives pour la textométrie des états des langues passées ?

Dans son introduction au deuxième atelier ATHIS, « L'historien, le texte et l'ordinateur », qui s'est tenu en novembre 2006 à Lyon, Jean-Philippe Genet a rappelé que les historiens – et en particulier les historiens des périodes anciennes – s'étaient plutôt détournés ces dernières années des méthodes informatiques de traitements de textes et de corpus. L'une des causes avancées en était le problème de l'apprentissage de ces méthodes face à leur importante sophistication, cette dernière étant tout autant liée à la complexification croissante des logiciels qu'à leur multiplicité, même si les plus connus, pour la France, restent sans doute Hyperbase, Lexico 3 et Weblex[1] – mais je reviendrai plus loin sur ce point. Une autre cause essentielle est l'adaptation de ces méthodes à des textes écrits dans des langues instables sur tous les plans (orthographe, syntaxe…), en particulier pour ce qui concerne les langues vernaculaires. Mais les communications qui ont suivi cette introduction ont présenté de nombreux apports dans divers domaines et, du fait même de leur richesse, ont finalement un peu masqué ces questions. Or, malgré les efforts faits par quelques-uns, ces problèmes doivent être résolus en priorité si l'on souhaite tout simplement que les historiens des périodes anciennes s'approprient ces méthodes et ces outils qui peuvent véritablement les faire avancer. En outre, cette appropriation raisonnée et réflexive pourrait nourrir les interrogations sur les enjeux beaucoup plus vastes qui concernent l'évolution même des conceptions de la connaissance, du fait des transformations du Web, par ailleurs largement abordées dans ce volume. Car il est essentiel que les historiens participent pleinement aux réflexions actuelles sur la notion de document numérique[2], par exemple, ou encore sur celle de l'information et de son traitement, au même titre que les scientifiques et les ingénieurs des autres disciplines.

[p. 158] Cette communication a pour ambition de poser des questions, sans prétendre apporter toutes les réponses. Le point de vue adopté est celui d'une historienne qui n'est ni linguiste ni informaticienne mais qui utilise depuis longtemps quelques outils bien rodés depuis une trentaine d'années et qui paraissent désormais pratiquement obsolètes pour certains spécialistes, alors même qu'ils n'ont jamais été réellement généralisés – il s'agit principalement de l'étude des concordances et des contextes d'une part et l'analyse factorielle par correspondance de l'autre. Et il faut souligner que jusque très récemment, ces deux types d'analyses étaient parmi les

[1] Ces logiciels ne sont pas disponibles en ligne, mais les sites mentionnés proposent de la documentation. Pour Hyperbase, http://www.unice.fr/bel/ spip.php?rubrique38 ; pour Lexico 3, http://www.cavi.univ-paris3.fr/Ilpga/ilpga/ tal/lexicoWWW/lexico3.htm ; pour Weblex, http://weblex.ens-lsh.fr/wlx/.

[2] Cf. R. T. Pédauque, *Le document à la lumière du numérique*, Paris, 2006.

rares à être traités par tous les logiciels[3]. C'est le point de vue d'une historienne qui aimerait profiter des avancées considérables effectuées ces dernières années mais sans pour autant y investir tout son temps (car la question du temps est bien sûr cruciale) ; qui souhaiterait enfin participer à la réflexion générale sur l'apport et la maîtrise de ces outils dans le cadre des enjeux évoqués ci-dessus. Le tout au risque de marteler quelques évidences.

Dans un premier temps, je m'attacherai à quelques réflexions autour du problème de l'apprentissage des outils informatiques, sachant que cet apprentissage est indissociable d'une réflexion épistémologique sur la nature de ces outils – liée en partie aux développements récents de la linguistique. Dans un second temps, je proposerai quelques types de traitements qui me paraissent utiles à l'historien, pour peu qu'il se les réapproprie et y réfléchisse, sans passer sous silence les problèmes d'adaptation aux langues anciennes.

La question de l'apprentissage

Un constat doit d'abord être effectué : l'apprentissage et l'adaptation des outils statistiques pour l'historien des périodes anciennes se heurtent en premier lieu à un foisonnement tout autant théorique que technique, lui-même déjà grand consommateur de temps. Pour préparer cette communication, je me suis plongée notamment dans les actes des Journées internationales d'Analyse des Données Textuelles, organisées depuis 1992, en ligne sur le site Lexicometrica[4]. Ces actes très abondants sont à première vue un peu effrayants pour l'historien qui n'a pas déjà une solide formation en la matière. Le simple dépouillement des résumés et des mots-clés des [p. 159] 104 communications des Journées de 2008 permet cependant de dégager quelques tendances. Il faut d'abord souligner une tendance négative qui nous concerne : les historiens sont très peu représentés, beaucoup moins en tout cas que les littéraires, les sociologues, les psychologues et même les chercheurs en technique marketing – sans même évoquer, bien sûr, les linguistes et les informaticiens qui dominent. Les rares communications portant sur des corpus médiévaux ou antiques, par exemple, sont avant tout le fait de linguistes ou de littéraires. Mais d'un point de vue plus positif, il m'a paru frappant – et très fécond – qu'au-delà des études de cas et des présentations d'applications ou d'amélioration de ces dernières, de nombreuses communications témoignent d'une véritable réflexion sur les relations entre objet étudié et logiciel utilisé. Et cette réflexion se retrouve dans bien d'autres ouvrages, par exemple dans les travaux du linguiste François Rastier[5] ou dans ceux du sociologue François Chateauraynaud (qui

[3] On pourrait ajouter à cette liste basique l'environnement thématique développé dans Hyperbase et la recherche des segments répétés dans Lexico 3.

[4] http://www.cavi.univ-paris3.fr/lexicometrica/jadt/index.htm. Les actes ne sont en ligne que depuis 1998.

[5] Voir notamment F. Rastier, *Arts et sciences du texte*, Paris, 2001.

est un des concepteurs du logiciel Prospero, sur lequel je reviendrai plus loin)[6]. Mais cette réflexion est pour l'heure un peu éparpillée ; elle mériterait donc, à mon sens, une véritable concertation qui permettrait de la rendre plus accessible.

Cet éparpillement se retrouve sur le plan même de la nomination de l'ensemble des méthodologies, en tout cas en français, qui permettent d'effectuer des analyses de textes ou de corpus assistés par l'informatique. En fait, il ne semble pas vraiment exister de consensus en la matière, car les noms mêmes dépendent de l'approche adoptée : lexicométrie, lexicologie, textométrie, logométrie, analyse textuelle des données… On trouve une pléthore de termes, mais qui ne renvoient jamais tout à fait à la même chose. Si le terme de lexicométrie paraît aujourd'hui peu employé, car il renvoie peut-être au concept un peu ancien de la simple association lexique/mesure, les autres sont utilisés plus régulièrement. La textométrie et l'analyse textuelle des données sont actuellement les termes les plus courants et témoignent d'un élargissement des perspectives. Le passage de la lexicologie à la textométrie renvoie en effet à un déplacement de l'approche du lexique à celui de l'appréhension d'un texte dans sa totalité. Quant à l'analyse textuelle des données, elle regroupe un ensemble d'analyses qui ne sont pas uniquement statistiques – c'est le terme qui a le sens le plus général. La logométrie, pour sa part, renvoie plutôt à l'appréhension du discours – ce qui conduit bien sûr à réfléchir au rapport entre texte et discours.

[p. 160] Ce déficit d'accessibilité est également lié à la question des cadres épistémologiques. La discipline qui entretient les liens les plus étroits avec ces méthodologies est sans doute la linguistique – ou plutôt certains courants de la linguistique. Or, ces liens entre linguistique et informatique sont, pour faire court, de plus en plus complexes, alors même que l'interaction est de plus en plus grande. En outre, il faut reconnaître que les historiens font souvent preuve d'une large méconnaissance des développements actuels de la linguistique. Il est vrai que l'on peut être rebuté par l'aspect parfois jargonnant de certains ouvrages de cette discipline, qui souffre de temps en temps d'un défaut de transmission – d'une overdose de technicité théorique, si l'on m'autorise ce rapprochement un peu paradoxal. Et l'on sait combien les historiens sont parfois rétifs à ce type de discours. Mais cela n'explique pas tout. Plus profondément, il existe au sein de la linguistique des courants qui divergent profondément dans leur approche, et les linguistes eux-mêmes conviennent parfois qu'il n'est pas toujours simple de s'y retrouver ou même d'établir des ponts entre eux. Pourtant, il est là encore indispensable de s'y plonger, car il est impossible de comprendre les logiciels que l'on souhaiterait s'approprier sans comprendre les postulats sur lesquels ils ont été construits.

---

[6] F. Chateauraynaud, *Prospero : une technologie littéraire pour les sciences humaines*, Paris, 2003.

Pour en revenir plus précisément à notre propos, les courants qui me sont apparus le plus fréquemment en relation avec les traitements informatiques sont la linguistique de corpus, la linguistique textuelle, la sémantique interprétative et l'analyse de discours, encore que cette dernière se situe plutôt à la frontière de la linguistique. Il est impossible dans le cadre de cette communication de présenter précisément tous ces courants, dont les dénominations sont tout de même relativement parlantes[7]. Je remarquerai simplement que, de plus en plus, certains chercheurs tentent de faire des liens entre ces courants, notamment en relation avec le traitement informatique, et que l'établissement de ces liens me semble être une véritable nécessité, même si la synthèse n'est pas forcément évidente comme l'a rappelé François Rastier dans sa communication intitulée « Que cachent les données textuelles ? » aux dernières Journées d'Analyse des Données Textuelles[8]. Ces tentatives me paraissent ouvrir des perspectives fécondes pour notre propre réflexion d'historiens, d'autant plus qu'elles rejoignent certaines de nos préoccupations et notamment toutes celles qui concernent les rapports entre [p. 161] les formes et les contenus d'un document. Je renvoie bien sûr à toutes les réflexions ouvertes dans le cadre de la fameuse « révolution documentaire », pour reprendre l'expression de Jacques Le Goff, qui se retrouvent actuellement dans de nombreuses disciplines historiques comme la codicologie, la paléographie, la diplomatique, etc. Plus généralement, une meilleure connaissance de ces divers courants peut enrichir nos propres discussions sur les notions mêmes de texte, de source, de document.

Du côté de l'informatique, l'éparpillement est sans doute plus important encore et peut aisément conduire le non-initié à se perdre dans ses méandres. D'une part, on se retrouve confronté à la complexité des langages informatiques, que ce soit dans le domaine de la programmation ou dans celui du codage, les deux étant bien sûr de fait indissociables. Dans le domaine du codage des métadonnées, il y a des efforts de standardisation, fondés sur la domination croissante du langage de balisage xml et les développements toujours plus importants de la TEI (Text Encoding Initiative) créée en 1987[9], ainsi que la mise en place de sa version simplifiée (et beaucoup plus accessible) en 1996, la TEI lite. Il faut également noter l'apparition de normes de codages plus ramassées comme le Dublin Core, créé en 1995 pour définir une description relativement compacte des ressources numériques[10]. Mais ces efforts de

---

[7] Pour une présentation orientée, mais synthétique et très fructueuse, voir Rastier, *Arts et sciences du texte*, *op. cit.* Il fournit par ailleurs une bibliographie.

[8] F. Rastier, *Que cachent les « données textuelles » ?*, *Actes des 9e Journées internationales d'Analyse Statistique des Données Textuelles*, Lyon, 2008, p. 13-26, p. 13 (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/index.htm).

[9] http://www.tei-c.org/index.xml.

[10] http://dublincore.org/. La norme Dublin Core comprend 15 champs, régulièrement retravaillés et qui peuvent faire l'objet de nombreux raffinements (titre ; créateur ; sujet ou mots-clés ; description ; éditeur ; contributeur ; date ; type de ressource ; format ; identifiant de la ressource ; source ; langue ; relation avec d'autres ressources ; couverture

standardisation ont provoqué un certain nombre de polémiques qui renvoient là encore à une réflexion épistémologique. On peut ici invoquer l'important débat qui s'est développé autour du « web sémantique » de Tim Berners-Lee et concrétisé par la conception de la RDF (Ressource Description Framework) en 1999, qui se veut basée, je le rappelle, sur un modèle conceptuel permettant de décrire n'importe quelle donnée. La critique principale est bien résumée par Rastier : selon lui, le web sémantique, je le cite, « entend remplacer le "web des documents" par le "web des données". En utilisant des ontologies, il s'agit de s'affranchir de la complexité des documents et de leur diversité linguistique et sémiotique… la donnée est alors conçue comme une simple chaîne de caractères »[11]. Il ne s'agit pas bien sûr, même pour Rastier, d'évacuer [p. 162] toute métadonnée ou même toute modélisation, mais au contraire d'articuler documents, données et métadonnées dans une perspective non réductrice.

D'autre part, je l'ai mentionné en introduction, les logiciels de traitement de données sont eux-mêmes nombreux. Chacun a développé ses propres codages, mais surtout ses propres traitements, ce qui fait que tous ces logiciels sont souvent complémentaires mais impossibles à articuler entre eux. C'est d'ailleurs la constatation des développeurs eux-mêmes, et je cite ici Serge Fleury, André Salem et Cédric Lamalle : « Les chaînes de traitement se présentent la plupart du temps comme des tunnels méthodologiques qui imposent que les corpus de textes soient mis dans une forme particulière pour pouvoir y pénétrer et qui restituent des résultats dans des formes qui leur sont propres interdisant pratiquement toute comparaison et toute analyse sur la pertinence et l'apport de chacune des étapes du traitement »[12]. Si l'on souhaite pouvoir s'approprier de manière point trop difficile les différents types de traitement proposés, il est très urgent de remédier à cet état de fait. C'est d'ailleurs tout l'objet de l'ANR Textométrie, qui regroupe les principaux concepteurs de logiciels d'analyse des données textuelles (en français), et dont l'objectif est de fédérer « les développements logiciels académiques du domaine pour mettre en place une plateforme modulaire en open-source » qui soit disponible en ligne ce qui, entre parenthèses, permettrait aussi d'en finir avec les incompatibilités liées aux systèmes d'exploitation et aux problèmes d'obsolescence des logiciels[13]. Une première version des logiciel a été mise en ligne en 2010[14].

---

spatiale ou temporelle ; droits).

[11] F. Rastier, *Que cachent les « données textuelles » ?*, art. cité, p. 13.

[12] C. Lamalle, S. Fleury et A. Salem, *Vers une description formelle des traitements textométriques*, *Actes des 8e Journées internationales d'Analyse Statistique des Données Textuelles*, Besançon, 2006, p. 583-593, p. 586.

[13] http://textometrie.ensolyon.fr/. Il fonctionne, mais tous les problèmes ne sout pas encore résolus.

[14] Stéphane Lamassé, Julien Alerini, Alain Dallo et Benjamin Deruelle ont également mis en ligne sur le site du Pireh un logiciel permettant d'effectuer différentes analyses statistiques (http://analyse.univ-paris1.fr). Voir la présentation de Stéphane Lamassé et Julien Alerini dans ce volume.

Appréhender et adapter les outils textométriques

Je voudrais dans un second temps énoncer quelques réflexions concernant l'appréhension et l'adaptation à notre discipline des outils textométriques, ainsi que quelques possibilités de traitements offertes. Deux grands éléments sont à considérer, mais qui interagissent bien sûr fortement.

[p. 163] Il y a d'une part, tout ce qui concerne la constitution et la préparation du corpus en vue d'un traitement statistique. Je ne m'étendrai pas ici sur la problématique du choix et de la définition du corpus, qui sont au cœur de débats importants dans le cadre de la linguistique de corpus, notamment au sein de la revue en ligne *Corpus*, mais qui, à mon sens, ne sont pas trop difficilement appréhendables par l'historien qui sait bien qu'un corpus est toujours orienté et que sa constitution doit faire l'objet d'une réflexion essentielle. Je renverrai cependant à la pertinente réflexion de Damon Mayaffre dans son article « Les corpus réflexifs : entre architextualité et hyper-textualité » paru dans la revue *Corpus* en 2002[15]. Sur la question de la préparation d'un corpus, plusieurs problèmes doivent être évoqués, particulièrement cruciaux pour les textes anciens. Deux hypothèses sont envisageables. On peut vouloir constituer un corpus à taille humaine, si l'on peut dire, c'est-à-dire un corpus fermé constitué en vue d'un travail de recherche précis, peu importe lequel en l'occurrence. On peut également décider de travailler sur un « grand corpus », éventuellement ouvert, avec possibilité de constituer des sous-corpus. C'est bien sûr une possibilité qui se développe de plus en plus, essentiellement par le biais d'Internet. Pour ce qui nous concerne, on renverra à la Base de Français Médiéval et plus généralement au Consortium pour le français médiéval créé en 2004, et qui pourrait constituer un exemple pour d'autres langues médiévales[16]. Dans les deux cas, certains problèmes sont similaires, et ce malgré la différence d'échelle.

Comment résoudre le problème de la variance des textes, sachant que, dans une culture manuscrite, il n'existe pas d'œuvre stable (le problème se pose peut-être moins pour les sources de la pratique, encore que…) ? Dans ce domaine, il y a sans doute des pistes très intéressantes du côté de la technique de l'alignement des textes. Cette méthode a été développée notamment par rapport à la question de traitement des corpus multilingues, mais aussi par rapport à celle du traitement des différentes versions d'œuvres littéraires contemporaines. Elle doit clairement être affinée pour les textes anciens, mais des recherches ont déjà été effectuées en la matière. En 2002, une équipe de linguistes suisses a par exemple montré qu'il pouvait être fructueux de construire une approche multicritères pour aligner des textes en français médiéval, fondée sur l'analyse des

---

[15] D. Mayaffre, *Les corpus réflexifs : entre architextualité et hypertextualité*, *Corpus* n°1 : *Corpus et recherches linguistiques*, 2002 (http://corpus.revues.org/ document11.html).

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

fréquences et des chaînes de caractère, sur celle [p. 164] des propriétés morphosyntaxiques et lexico-sémantiques, mais aussi sur le plan de la structure typographique et rhétorique[17].

Toujours dans le cadre de la culture manuscrite, comment prendre en compte la matérialité de la source ? Là encore, il y a des possibilités mais qui ne sont pas encore réellement développées, en tout cas dans l'optique d'un traitement statistique. Quelques projets très intéressants ont été mis en place, mais ils ont été limités pour l'instant à l'édition électronique. On peut citer par exemple l'édition, sous la direction d'Olivier Cullin, du *Graduel de Bellay* sur le site de l'École nationale des chartes, qui constitue un essai de représentation de toutes les facettes du manuscrit par le biais du multimédia[18]. Cette question doit donc, à mon sens, être reliée à la question du multimédia. Il a parfois été noté, mais cela paraît une évidence pour le médiéviste, que le manuscrit enluminé et rubriqué était l'ancêtre par excellence du document multimédia d'aujourd'hui. Selon François Rastier, par exemple, « on sait que toute performance linguistique peut mettre en jeu plusieurs sémiotiques (prosodiques, kinésiques, typographiques, par ex.), mais la numérisation permet de les combiner dans des textes multimédias, qu'annonçaient d'une manière peut-être indépassable les manuscrits enluminés à figures »[19]. Encore faudrait-il que cette remarque ne se transforme pas en cliché. Un traitement informatique raisonné des manuscrits médiévaux rubriqués et/ou enluminés me paraît une perspective prometteuse, non seulement pour l'histoire du manuscrit en général, mais aussi dans le cadre d'une analyse textuelle ; car tout historien des manuscrits sait bien que les différents éléments de la page manuscrite – texte nu, commentaire, rubrication et images – fonctionnent ensemble et s'éclairent les uns les autres.

Un dernier problème, peut-être le plus important pour l'historien travaillant sur des états de langue anciens, est celui de la lemmatisation. Apparemment, cette dernière ne fait plus tellement débat pour les linguistes qui travaillent sur les langues contemporaines, dans la mesure où la lemmatisation par annotation est relativement simple avec un appareillage morpho-syntaxique. La dernière version d'Hyperbase a d'ailleurs intégré une interface avec le logiciel Cordial qui permet, pour les langues contemporaines, d'annoter les lemmes et les catégories grammaticales[20]. Mais pour les langues anciennes, le [p. 165] traitement est beaucoup plus problématique. Il s'agit pourtant d'un chantier important car les perspectives sont réelles pour tenter d'automatiser au moins en partie la lemmatisation des textes médiévaux. On peut invoquer l'exemple du projet

---

17 H. Ghorbel, G. Coray, A. Linden, O. Collet et W. Azzam, *L'alignement multicritères des textes médiévaux*, *Lexicometrica : Alignement textuel dans les corpus multilingues*, 2002 (http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm).

18 http://bellelay.enc.sorbonne.fr/presentation.php.

19 Rastier, *Arts et sciences du texte*, *op. cit.*, p. 81-82.

20 Cordial est à l'origine un correcteur orthographique et grammatical, mais les concepteurs ont ensuite développé le logiciel Cordial Analyseur qui permet l'étiquetage morpho-syntaxique. C'est un logiciel commercial (http://www.synapse-fr.com/CordialoAnalyseur/PresentationoCordialoAnalyseur.htm).

« Manuscript » sur le russe ancien, lequel combine la triple difficulté d'une graphie différente, d'une langue à déclinaison et de l'instabilité de cette dernière[21]. L'objectif du projet a donc été de créer, à partir d'un vaste corpus de textes en russe ancien et des dictionnaires existants, un analyseur morphologique automatique permettant de reconstruire les paradigmes des lemmes, en tenant compte à la fois des variantes orthographiques, des déclinaisons et des conjugaisons. À partir de là, il est possible d'effectuer des lemmatisations. Le logiciel, qui en est à sa quatrième version[22], est disponible en ligne pour interrogation, mais non pour ajouter des donne´ es directement.

Évidemment, tout cela est lié à la question plus générale des métadonnées et de l'annotation des corpus, dont les possibilités vont bien au-delà de ces trois problèmes particuliers, puisque les langages informatiques actuels ont permis la multiplication des formes de balisages. Mais là se posent cruellement, non seulement la question du choix du type de codage, mais aussi les questions de l'investissement en temps et de l'adaptation. Ces interrogations sont anciennes et ont précédé l'apparition du xml, mais elles apparaissent plus que jamais d'actualité. La TEI, par exemple, ou même sa version allégée, est-elle vraiment adaptable aux textes anciens ? De fait, plusieurs pistes sont actuellement explorées et la question de l'annotation des corpus apparaît de manière récurrente dans les réflexions des chercheurs en analyse de données textuelles. Certains réfléchissent à la mise en place d'outils complets et facilement adaptables. On peut citer par exemple les recherches de Jean-Marie Viprey, de Virginie Léthier et de leurs équipes franc-comtoises pour la constitution d'un environnement d'annotation intitulé DiaTag (pour *dialogic tagging*), combinant la plus grande automatisation possible à un « contrôle humain » rigoureux, et je cite ses concepteurs :

> « Le principe général de DiaTag est de faire alterner des phases automatiques et des phases de dialogue où les situations non univoques sont soumises à une décision humaine, de la manière la [p. 166] plus ergonomique possible. Un autre principe majeur est de permettre l'enrichissement et l'édition progressifs des ressources, notamment des dictionnaires "livrés" avec le système, dans le cours même des opérations d'annotation »[23].

Un effort particulier a été fait pour traiter les formes composées, plus complexes. L'intérêt de cette démarche est bien sûr qu'elle est contributive.

Il faut cependant souligner que ces recherches ne concernent pour l'instant que les langues naturelles contemporaines ; en outre, elles renvoient au fait que la constitution de modèles

---

[21] V. A. Baranov, A. N. Mironov, A. N. Lapin, I. S. Melnikova et A. A. Sokolova, *Development of the processing and visualization technologies for the linguistic information in the manuscript system : lemmatization*, *Actes des 9e Journées inter- nationales d'Analyse Statistique des Données Textuelles*, Lyon, 2008, p. 137-145 (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/index.htm).

[22] http://manuscripts.ru/indexoen.html. Il y a une présentation en anglais.

[23] J.-M. Viprey et V. Léthier, *Annotation linguistique de corpus : vers l'exhaustivité par la convivialité*, *Actes des 9e Journées internationales d'Analyse Statistique des Données Textuelles*, Lyon, 2008, p. 1151-1161, p. 1153 (http:// www.cavi.univ-paris3.fr/lexicometrica/jadt/index.htm).

d'annotation ne résout pas tout, ce qui rejoint en partie la critique de certains linguistes sur le web sémantique. D'autant qu'il faut être attentif à la manière dont sont conçus les schémas définissant les éléments de balisages et leurs attributs. Il faut en effet suffisamment de souplesse pour qu'ils soient adaptables à nos objectifs de recherche. Dans le domaine des langues anciennes, une piste intéressante est à cet égard celle du projet Khârtès, qui porte sur un projet d'édition et d'étude linguistique des chartes françaises de Wallonie. Les initiateurs de ce projet ont conçu un système d'annotation spécifiquement destiné aux chartes, très léger mais facilement manipulable, et qui serait sans doute adaptable à d'autres chartes et peut-être à d'autres types d'actes[24]. Enfin, il faut songer à implémenter ces annotations dans les logiciels : c'est déjà en partie le cas pour le logiciel Weblex et c'est un des aspects importants du projet Textométrie (on peut ici renvoyer aux réflexions de Bénédicte Pincemin)[25]. Le chantier qui s'est ouvert en matière de modélisation des données et des métadonnées est donc très vaste et dépasse largement les enjeux en matière d'histoire médiévale (ou ancienne). Mais je pense que l'historien des périodes anciennes a ici un rôle à jouer, du fait même de la nature des sources sur lesquelles il travaille et qui invite à une certaine souplesse et à une adaptation continue.

Les remarques précédentes valent pour tous les types de corpus. En revanche, si de nouvelles perspectives se sont ouvertes avec la mise en ligne de très grands corpus, ces derniers induisent toute une série de nouveaux problèmes. Il y a d'une part le fait que la plupart [p. 167] des corpus accessibles pour l'instant le sont en html ou en pdf. Se pose donc le problème du dépôt des données et de l'extraction. Si je souhaitais, par exemple, travailler sur un très grand corpus en moyen anglais, celui du projet TEAMS[26], je devrais faire face à un long travail préparatoire… du simple fait que les éditions sont en html et qu'il faut commencer par nettoyer les textes de tout l'apparat critique, car ce sont des éditions récentes et annotées (ce qui est bien sûr très précieux !). Il existe des solutions, dans la mesure où il est possible de transformer une page html en page xml à l'aide de feuilles de style[27], mais c'est une solution relativement lourde à mettre en œuvre. L'idéal serait de trouver une solution légère d'extraction automatique. Il est donc impossible de penser la question du traitement des grands corpus sans la mettre en lien avec la problématique de l'édition électronique. Mais là encore, des réflexions vont dans ce sens, par exemple au sein de l'équipe canadienne dirigée par François Daoust (qui est par ailleurs le

---

[24] N. Maziotta, *Le texte dans tous ses états. Philosophie d'encodage du projet Khartês*, Actes des 7e Journées internationales d'Analyse Statistique des Données Textuelles, Louvain, 2004, p. 793-803 (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/tocJADT2004.htm).

[25] B. Pincemin, *Modélisation textométrique des textes*, *Actes des 9e Journées internationales d'Analyse Statistique des Données Textuelles*, Lyon, 2008, p. 949-960 (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/index.htm).

[26] http://www.lib.rochester.edu/camelot/teams/tmsmenu.htm.

[27] Cela est notamment possible avec le logiciel jEdit (http://www.jedit.org/).

concepteur du logiciel SATO)[28].

En ce qui concerne les traitements statistiques proprement dits, la question est de savoir quel traitement utiliser pour quel objectif. Et en tant qu'utilisatrice lambda, si l'on peut dire, qui considère de plus que le traitement informatique des données n'est pas l'unique méthode de travail à exploiter, je suis bien consciente que je n'utilise pas, et de loin, toutes les possibilités offertes par les logiciels, ni même d'ailleurs, par un seul logiciel. Si l'on veut convaincre davantage d'historiens des mérites de ces traitements, un travail pédagogique constitue un préliminaire indispensable car la diversité et la complexité de ces traitements sont pour le moins exponentielles. D'où aussi, il faut le répéter, la nécessité d'un environnement intégré et accessible – en ligne donc – qui peut véritablement ouvrir des perspectives si tant est qu'il soit conçu dans un esprit d'accessibilité. Mais venons-en aux traitements proprement dit. Actuellement – et sans même parler de toutes les possibilités offertes par des corpus proprement annotés – plusieurs directions importantes ont été prises, dont les historiens pourraient, me semble-t-il, tirer grand profit. Étant donné mes propres axes de recherche, les directions que je mention-[p. 168]nerai concernent surtout les possibilités d'analyses du contenu des textes. Ce qui ne veut pas dire qu'il faut négliger la forme des textes et de la langue, même si l'impression dominante est qu'elle est pour l'heure davantage exploitée par les littéraires (je renvoie à nouveau à tous les travaux effectués sur la Base de Français Médiéval).

Il faut d'abord signaler les nombreux travaux sur les cooccurrences et leurs cousines anglo-saxonnes, les collocations, qui se sont beaucoup renouvelés et qui conduisent à la constitution de réseaux lexicaux – ce qui est extrêmement précieux si on veut faire un peu de sémantique. La notion de cooccurrence en lexicologie n'est certes pas nouvelle, mais les approches ont beaucoup évolué. Dans les quelques réflexions qui suivent, je m'inspire en grande partie d'une communication de Damon Mayaffre qui a présenté très clairement les enjeux liés à cette notion. Pour lui, la statistique co-occurrentielle permet en effet de passer « d'une approche formelle, nucléaire ou positiviste du corpus à une approche contextualisante c'est-à-dire déjà sémantique. Avec la cooccurrence, la statistique textuelle met un pied dans une sémantique de corpus qui lui était jusqu'ici interdite et réaffirme par là sa vocation herméneutique »[29]. D'autant qu'il ne s'agit plus seulement d'étudier des couples de mots au sein d'un corpus, d'un texte ou d'une unité textuelle, mais bien de construire de véritables réseaux sémantiques en constituant des pôles, des

[28] F. Daoust, J. Duchastel, Y. Marcoux et E. Rizkallah, *Pour un modèle de dépôt de données adapté à la constitution de corpus de recherche*, *Actes des 9e Journées internationales d'Analyse Statistique des Données Textuelles*, Lyon, 2008, p. 355-367 (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/index.htm). Pour le logiciel SATO, voir la présentation en ligne (http://www.ling.uqam.ca/sato/outils/ sato.htm).

[29] D. Mayaffre, *Quand « travail », « famille », « patrie » co-occurrent dans le discours de Nicolas Sarkozy. Étude de cas et réflexion théorique sur la co-occurrence*, *Actes des 9e Journées internationales d'Analyse Statistique des Données Textuelles*, Lyon, 2008, p. 811-822, p. 812 (http://www.cavi.univ-paris3.fr/ lexicometrica/jadt/index.htm).

« nœuds du tissu textuel » qui s'articuleront les uns aux autres. Mayaffre a montré concrètement l'importance de cette approche en partant du terme « travail » dans les discours de campagne de Nicolas Sarkozy, qui forme le nœud d'un réseau incluant également la famille et la patrie. C'est Weblex qui, le premier, a permis de construire des lexicogrammes représentant de tels réseaux. Mais Lexico 3 s'est récemment adjoint les services d'un module intitulé Coocs, développé par William Martinez, qui permet d'arriver au même résultat[30]. Et dans sa dernière version, Hyperbase a également introduit ces lexicogrammes.

Une autre direction, mais complémentaire de la première, a été prise par les travaux sur la topologie textuelle qui s'est beaucoup développée ces dernières années et qui est particulièrement intéressante pour comparer des séries chronologiques ou plusieurs états d'un même texte entre eux. L'idée est en effet de considérer l'organisation topographique des textes et des corpus, de réaliser en quelque [p. 169] sorte leur cartographie. D'avoir la possibilité de voir, au sens propre du terme, se déployer une occurrence, une cooccurrence ou un segment répété dans l'espace textuel considéré. Dans Lexico 3, par exemple, « le corpus est représenté, à l'écran, phrase à phrase ou paragraphe après paragraphe, par autant de carrés successifs. Et ces carrés se colorent ou restent vierges selon la présence ou non de la forme linguistique recherchée. L'outil permet donc de localiser et visualiser des formes dans la suite continue du corpus »[31]. Cette notion de topologie textuelle est liée à la notion de rafale chère à Pierre Lafon, qui renvoie à la surutilisation d'un mot et qui peut être définie mathématiquement, mais aussi topologiquement[32]. Mais elle est également tout à fait complémentaire de celle des cooccurrences. Il s'agit bien d'envisager les textes sous tous les angles possibles. Dans le même ordre d'idées, on peut évoquer la notion de résonance textuelle développée récemment par André Salem, qui combine la topographie, l'alignement et la fréquence pour « étudier simultanément des textes dont chacun entretient des rapports étroits avec l'autre du point de vue de sa structuration »[33]. Cette notion est intéressante pour étudier les variantes d'une œuvre ou les corpus multilingues, mais on pourrait imaginer d'autres applications, comme l'étude de séries de textes apparemment très homogènes (chartes, préambules, prologues…).

Pour terminer, je voudrais évoquer non plus un type de traitement statistique, mais une

---

30 http://coocs.eu.md/page.php.

31 D. Mayaffre, *L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie/ topologie textuelle (partie I)*, *Lexicometrica : Topographie et topologie textuelle*, 2007 (http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9. htm).

32 P. Lafon, *Dépouillements et statistiques en lexicométrie*, Paris et Genève, 1984. Voir aussi E. Brunet, « Navigation dans les rafales », *Actes des 8e Journées internationales d'Analyse Statistique des Données Textuelles*, Besançon, 2006, p. 15-29 (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/tocJADT 2006.htm).

33 A. Salem, *Introduction à la résonance textuelle*, *Actes des 7e Journées internationales d'analyse statistique des données textuelle*s, Louvain, 2004, p. 986-992, p. 986
(http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/tocJADT 2004.htm).

approche un peu différente de l'analyse textuelle, incarnée par le logiciel Prospero, qui relève de ce que Philippe Cibois a appelé la « mouvance documentaire ». Prospero a été conçu par des sociologues pour « permettre la description et l'analyse des dossiers complexes, marqués par de longues séries de textes et de discours hétérogènes »[34]. Et ce de manière collective, [p. 170] puisque si le codage préliminaire est pratiquement nul, l'utilisateur est constamment sollicité pour redéfinir les grandes familles d'objets ou d'entités de Prospero (les personnages, les catégories et les collections), ainsi que les différents traitements possibles sur ces entités. Il y a là une véritable réflexivité sur l'outil employé et une véritable interaction. L'approche est très différente de celle des autres logiciels et j'avoue ne pas être certaine que ce logiciel puisse être adapté aux corpus en langues anciennes. Mais les réflexions de la communauté d'utilisateurs de Prospero sont de toute manière intéressantes pour leurs aspects théoriques, en particulier sur la question de l'interprétation des données.

En conclusion, les perspectives offertes sont énormes, et pas seulement au niveau technique. J'espère avoir montré, en effet, comment la croissance exponentielle des possibilités offertes en informatique s'accompagnait forcément d'une réflexion intense sur les enjeux épistémologiques ouverts pour notre discipline et plus généralement pour toutes les disciplines en sciences humaines. Mais ces perspectives se heurtent clairement à un problème d'apprentissage et à un problème d'investissement. Si ces problèmes ne sont pas résolus, non seulement les historiens n'auront pas accès à ces méthodes, mais en plus ils ne pourront contribuer à une réflexion générale sur les changements de paradigmes actuels en matière de traitement des connaissances.

[34] Association Doxa, *Prospéro et l'analyse des dossiers complexes*, http://prospero.dyndns.org:9673/prospero/accesopublic/02otextesosuroprospero/01o prosperoo2002.