Table ronde entre M. Bigey, A. Guerreau et V. Lethier (animée par T. Rainsford)

La discussion a été menée autour de trois thématiques : les enjeux de la numération, les modalités du traitement et enfin les enjeux liés à la discipline.

V. Lethier et **M. Bigey** ont tout numérisé, alors qu'**A. Guereau** travaille sur des corpus déjà numérisés. **M. Bigey** a consacré entre 3 et 4 jours au traitement de chaque roman (numérisation en une heure, puis relectures). **V. Lethier** évoque également les contraintes liées aux normes de conservation : pas de chaleur, pas de lumière intensive, besoin d'un scanner qui respecte la taille du quotidien, nécessité de ne pas désarticuler les volumes. Pour 1 millions de mots, elle estime à 22 heures le temps passé à scanner et à enregistré. Le travail d'OCRisation représente au bas mot 180 heures pour 1 million de mots, l'annotation linguistique contrôlée un volume de 13-15h par million de mots. Les typographies sont difficilement reconnues et l'espacement entre caractères est systématiquement erroné, ce qui a nécessité la restauration en amont du fichier image et d'apprendre des gabarits au logiciel. **A. Guerreau** recommande pour le passage du texte imprimé au texte numérisé de consacrer un maximum de temps au scan lui-même : un bon scan permettant d'économiser beaucoup de temps. Il faut aussi bien faire attention au réglage de son scanner : luminosité, contraste, et vérifier à quoi ressemblent les caractères pour ne pas trop perdre au scan. Il a lui-même testé deux types de logiciel OCR : FineReader qui passe par la reconnaissance de squelettes couplé avec une procédure dictionnaire. Le stock de mots possibles qu'utilise le dictionnaire peut être modifiable : il faut vérifier si le logiciel accepte un dictionnaire personnel. Il a également utilisé le logiciel Gamera (applications d'analyse et de reconnaissance de documents), qui travaille strictement en mode image.

V. Lethier pense qu'il est essentiel de ressortir du mode linéaire de correction. Elle recommande DIATAG, un logiciel qui permet des phases de dialogue avec l'expert et donc de trancher quand l'outil informatique n'en est pas capable : cet outil permet également un enrichissement progressif des ressources. Elle préfère pour sa part travailler avec du semi-automatique : pour conserver la possibilité de trancher en cas de doute. Par ailleurs, elle souligne l'importance du temps consacré à la constituion des corpus et pose la question de leur devenir. Après le travail de structuration obligatoire, il est nécessaire de suivre les métadonnées : les compétences qui ont présidé à l'élaboration du corpus ne sont pas forcément celles des littéraires ou des linguistes : il lui paraît essentiel de réinterroger le processus de formation, de se demander dans quelle mesure ces activités peuvent être valorisées et quelle valeur scientifique il faut attribuer à ces corpus et aux efforts fournis.

A. Guerreau insiste quant à lui sur le caractère essentiel de la mise en ligne en *open access*. Il constate qu'en France peu de bibliothèques fonctionnent en open source. Il se souvient de l'arrivée laborieuse de Koha (système intégré de gestion des bibliothèques) dans les bibliothèques.

V. Lethier pose la question du partage de ces éléments. Le dictionnaire du *Petit Comtois* est trié par catégories grammaticales et comprend des éléments de géolecte. **A. Guerreau** insiste pour dire qu'on ne doit pas perdre de vue que les chercheurs ont tout intérêt à ce que leurs idées se répandent. Pour lui, la question des droits sur les textes médiévaux se pose surtout en termes de sociologie du milieu des intellectuels : tous les éditeurs font des menaces, qui se révèlent efficaces, mais qui ne sont pas légitimes en droit.

Concernant l'utilisation des logiciels, **M. Bigey** aimerait un logiciel qui réunisse toutes les fonctionnalités qu'elle utilise. Elle trouve le dictionnaire électronique spécifique extrêmement

efficace. L'analyse factorielle des correspondances permet de prendre immédiatement conscience de pôles et de voir quels textes sont par rapport au lexique total en sous- ou en sur-représentation. **V. Lethier** souhaite se mettre très prochainement à TXM. Elle utilise SATO (UQAM), qui permet de procéder à des annotations et de les exporter, ainsi que Hyperbase qui permet de faire apparaître des cooccurrences et de créer des représentations arborées. Elle est très satisfaite notamment de la facilité de navigation entre les indices et le texte. Elle se sert également du logiciel Astartex, téléchargeable, ainsi que Lexico 3 mais ce dernier logiciel ne répond pas à ses attentes au niveau du retour au texte. Selon elle, ICTeNA (Université de Franche-Comté) est particulièrement adapté pour naviguer sur des petits corpus mais il est dépourvu de tout indice quantificatif et ne permet donc pas de comparer les différentes partitions. **A. Guerreau** utilise pour sa part PhiloLogic, un logiciel développé à Chicago qui permet de manipuler des corpus de grosse taille, mais avec un mot et rien de plus ; on peut indexer le corpus avec une vingtaine de champs. Il est passé à l'automne 2011 à TXM : il peut saisir autant d'informations qu'il le souhaite pour chaque texte et chaque élément. Il ne voit pas pour l'instant quelles sont les limites de ce logiciel.

Y.-F. Le Lay ajoute que l'un des intérêts R et de TXM, ce sont les questions d'export des tables et surtout des figures. Bien souvent on se retrouve avec une image dont on ne sait que faire et qui est impubliable. Le format vectoriel est donc très appréciable et les problèmes de superposition peuvent être réglées dans Inkscape ou Illustrator. Ce sont des choses très partageables. Le défaut de TXM et de Lexico à son sens, c'est que bien souvent on ne garde que l'exposant, une donnée qui ne parle qu'au spécialiste d'analyse textuelle. Si l'on veut parler avec les autres disciplines, il faut convertir, or peu de gens le font. Il insiste sur la nécessité de partage et de communication dans un langage parlé par le plus grand nombre.

A propos de l'export dans d'autres logiciels, **M. Bigey** utilise le logiciel NooJ, un logiciel en *open source*, facile d'accès et collaboratif.

M. Bigey et **V. Lethier** insistent toutes deux sur le fait que rester uniquement sur les nombres n'a aucun sens. Il faut éviter l'écueil qui consiste à partir dans des stéréotypes trop risqués. Cela passe par une exploration systématique du co-texte, comme unité supplémentaire, essentielle. **A. Guerreau** rappelle qu'il faut impérativement éviter de s'imaginer que tout le monde comprend un graphique factoriel, qui en soi ne prouve rien : il s'agit seulement d'un outil exploratoire pour effectivement comprendre quelque chose. Beaucoup de médiévistes ne veulent pas encore entendre parler de ces méthodes : ils constatent certes que ces nouvelles approches fonctionnent mais ne veulent pas admettre la pertinence de résultats obtenus par des méthodes auxquelles ils ne comprennent rien.

V. Lethier évoque que le travail d'annotation va de pair avec une profonde interrogation sur la pertinence de certaines catégories grammaticales figées. Cela pose la question de savoir où s'arrête le texte et des unités à prendre en compte. On a un peu tendance à être sensible au vertige de la fréquence, mais à l'inverse, certains hapax font vraiment sens.