Lexicométrie et corpus multilingues

Table-ronde

Abstract

Beyond automatic parallel text alignment, which is now well-known of our scientific community, this panel session focuses on how to *extend statistical techniques* in view of *exploring multilingual textual data*. As regards *parallel corpora*, new tools and methodologies have emerged. Processing *comparable corpora* (i.e. made-up of similar texts which are not the translation of one another) is also a significant challenge. Textual statistics for monolingual corpora can be adapted to this new type of data. Furthermore, some corpora are written in *languages* which raise new issues as concerns textual statistics softwares: for example the management of the characters encoding, the corpus tokenisation into sensible word-like units, or the definition of clear and coherent linguistic annotation schemes. International standards have recently been published and others are in preparation. They constitute efficient guidelines for corpus and linguistical ressources encoding. As they deal with the genuine diversity of languages throughout the world, these standards allow the comparability and the reusability of textual data.

Résumé

Par delà l'alignement automatique de corpus, fondamental mais bénéficiant déjà d'une certaine maturité et d'une bonne notoriété, cette table-ronde s'intéresse à l'*extension* des *techniques et applications lexicométriques* en contexte multilingue. Sont bien évidemment concernés les *corpus parallèles*, pour l'analyse et l'exploitation desquels peuvent être proposés de nouveaux outils et méthodes. Sont également en jeux les *corpus comparables* (non alignés), qu'il s'agit d'être capable d'exploiter moyennent l'adaptation de procédures d'analyse statistique jusque là pratiquées sur des corpus monolingues. Sont enfin également considérés les corpus dont la *langue* introduit de nouvelles questions théoriques et techniques pour les logiciels de lexicométrie, par exemple pour le codage des caractères, le découpage en « mots », ou le codage des informations morphosyntaxiques. Des *normes internationales* récentes et en cours d'élaboration guident maintenant le codage des corpus et des ressources linguistiques de tous ordres. Prenant en compte la diversité des langues, elles visent à favoriser la comparabilité et la réutilisabilité des données textuelles.

Mots-clés : corpus parallèles, corpus comparables, corpus en langues orientales, statistique textuelle, spécificités, réseaux de cooccurrences, segments répétés, représentation topographique de textes parallèles, interfaces de navigation, classification automatique, analyse canonique non linéaire, lexiques multilingues, jeux de caractères, découpage en mots, étiquetage linguistique, normalisation.

1. Présentation du thème

Le besoin se fait de plus en plus sentir d'outils appropriés à la consultation et à l'analyse de corpus multilingues, alignés ou non. Complémentairement aux outils d'alignement (utiles pour constituer les corpus) et d'appariement de termes, il s'agit d'outiller des modes de parcours, de synthèse, et d'extractions contextuels, offrant une approche moins purement terminologique et plus textuelle de ces données multilingues.

C'est une perspective de développement particulièrement importante pour les outils actuels de statistique textuelle. En effet, le traitement des corpus multilingues dans un cadre jusqu'à présent monolingue conduit soit à mêler les textes des différentes langues dans les calculs (sauf un calcul de contraste en les définissant comme des parties), soit à les traiter de façon totalement indépendante, aucune de ces deux voies n'étant pleinement satisfaisante. Il existe bien

quelques bons concordanciers pour corpus parallèles, mais ceux-ci n'offrent pas toute la richesse des calculs statistiques mis au point en lexicométrie.

Différents chantiers sont à considérer :

- la prise en compte de jeux de caractères qui sortent de l'ASCII ou de l'isoLatin1 (arabe, chinois,...)
- la prise en compte de modes de segmentation relatifs aux langues : la notion de forme délimitée par deux caractères séparateurs comme approximation du mot n'est pas directement transposable, cf. par exemple le cas des mots composés en allemand, ou des écritures par idéogrammes syllabiques.
- l'adaptation des calculs statistiques et des interfaces aux calculs alignés : traitements sur une langue et affichage des (ou accès aux) correspondances dans l'autre langue ; interrogations croisées ou/et alternées sur les différentes langues ; rôles bien différenciés des différentes langues (par ex., ce qui éclaire le mieux un calcul de concordance sur une langue n'est pas nécessairement un semblable calcul de concordance sur l'autre langue).
- la conception de nouvelles procédures et calculs statistiques, plus particulièrement liés aux besoins d'investigation que peuvent avoir des utilisateurs (enseignants de langues, linguistes, traducteurs, industriels, professionnels de l'information) sur ce type de corpus.

Le but de cette table-ronde est de réunir les personnes de la communauté des statistiques textuelles intéressées par ces questions, afin :

- de débattre sur les perspectives les plus importantes et intéressantes d'extension des techniques lexicométriques à ce terrain des corpus multilingues
- et éventuellement d'organiser les efforts de recherche et de développement, cette rencontre favorisant la prise de contacts pour qu'ensuite se positionnent et se coordonnent les acteurs intéressés.

2. Résumés des interventions

Coordination de la table-ronde : ***Bénédicte Pincemin***

CNRS, LLI – Université Paris 13 – Avenue J.-B. Clément – 93430 Villetaneuse – France
http://www-lli.univ-paris13.fr/ benedicte.pincemin@centraliens.net

2.1. *Mathieu Valette*

INALCO, Centre de Recherche en Ingénierie Multilingue – 2, rue de Lille – 75343 Paris cedex 7 – France
http://crim.inalco.fr/ mvalette@inalco.fr

Encore récente, la linguistique de corpus ouvre de nouveaux champs de prospection tant en matière de recherche (analyse de données attestées) que d'application (ingénierie du document). Mais parce qu'elle repose sur la technologie informatique, une minorité de langues en bénéficient réellement : celles écrites avec l'alphabet latin.

Toutefois, l'encodage d'un grand nombre d'alphabets et la normalisation de la structuration des textes (Unicode, TEI) permet d'envisager le traitement automatique de la plupart des langues écrites et subséquemment, le renouvellement de la problématique comparatiste historiquement définitoire des sciences du langage.

L'Institut National des Langues et Civilisations Orientales, et en son sein le Centre de Recherche en Ingénierie Multilingue, sont particulièrement désireux de participer à la pro-

blématisation de ce champ encore peu exploré. Car l'enjeu n'est pas exclusivement théorique et scientifique, il concerne également l'accès vital des langues orientales au secteur quaternaire (ou « nouvelle économie »).

2.2. *Maria Zimina*

SYLED EA 2290 – Université de la Sorbonne nouvelle Paris 3 – 19, rue des Bernardins – 75005 Paris – France
http://www.cavi.univ-paris3.fr/ilpga/ED/student/stmz/ zimina@msh-paris.fr

L'utilisation des pratiques de *textométrie* pour le traitement automatique de corpus parallèles constitue une piste de recherche prometteuse. Dans le contexte multilingue, les méthodes quantitatives (telles que la *classification automatique*, les *spécificités*, le *calcul des réseaux de cooccurrences*, etc.) apportent un éclairage précieux sur des régularités dans la structuration de discours parallèles. Ces méthodes peuvent être utilisées pour l'appariement automatique des vocabulaires et des fragments de texte. Elles reposent entièrement sur des ressources construites à partir de corpus parallèles eux-mêmes. Les fréquences et les distributions d'unités textuelles (formes, segments répétés, etc.) au sein de chacun des volets multilingues servent de points de repère pour l'identification et l'extraction des correspondances. Appuyée sur l'utilisation de la *représentation topographique* du *bi-texte*, l'approche textométrique permet d'envisager la conception de nouvelles procédures informatiques liées aux besoins d'investigation que peuvent avoir des utilisateurs (linguistes, traducteurs, professionnels de l'information, etc.) sur ce type de corpus.

Martinez W. et Zimina M. (2002). Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues. In *Actes des JADT 2002*.

Zimina M. (2000). Alignement de textes bilingues par classification ascendante hiérarchique. In *Actes des JADT 2000*.

Zimina M. (2002). Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues. In Véronis J. (Ed.), *Lexicometrica*, n° spécial Corpus alignés : http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm

2.3. *Cyril Goutte*

Xerox Research Center Europe – 6 ch. de Maupertuis – 38240 Meylan – France
http://www.xrce.xerox.com/ Cyril.Goutte@xrce.xerox.com

Pour de nombreuses applications multilingues (par exemple la traduction automatique) il est nécessaire de disposer d'un corpus aligné ou parallèle, typiquement issu de traductions. Dans d'autres domaines, il est difficile de disposer de tels corpus – il faut alors travailler à partir de corpus comparables (Déjean *et al.*, 2002), traitant du même sujet sans pour autant avoir de correpondance un-à-un. Dans ce cadre, on évoquera l'utilisation de techniques d'analyse des données qui sont le prolongement des décompositions spectrales types « Latent Semantic Analysis » utilisées en monolingue. En particulier, on présentera l'utilisation de l'analyse canonique, et son extension au domaine non-linéaire (Lai et Fyfe, 2000), pour l'enrichissement de lexiques bilingues à partir de corpus comparables.

Déjean, Gaussier et Sadat (2002). Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. In *Proceedings of COLING'2002*.

Lai et Fyfe (2000). Kernel and nonlinear canonical correlation analysis. *IJNS*, vol. (10/5) : 365-377.

2.4. *Laurent Romary*

LORIA UMR 7503 – Campus Scientifique – BP 239 – 54506 Vandœuvre-lès-Nancy CEDEX – France
http://www.loria.fr/equipes/led Laurent.Romary@loria.fr

Perspectives de normalisation pour les corpus et lexiques multilingues

Afin d'effectuer des recherches de plus en plus fines sur la langue, il est nécessaire de disposer d'une infrastructure stable mais flexible pour représenter différents types de ressources linguistiques : corpus annotés et lexiques. Au delà de la simple représentation de données primaires, pour lesquelles le cadre de la TEI (Text Encoding Initiative) est maintenant bien reconnu, je présenterai les grandes lignes d'action du nouveau comité TC 37/SC 4 de l'ISO qui travaille d'une part sur la modélisation d'annotations multi-niveaux et de ressources linguistiques multilingues. Je présenterai en particulier le projet de mise en œuvre d'un registre de catégories de données en linguistique qui, dans une perspective internationale, devrait permettre de disposer d'un référentiel de comparaison de différents schémas d'annotation ou de bases lexicales.

Je finirai en montrant comment de tels standards devraient nous permettre de mieux échanger des données et des outils au sein des communautés de linguistique et de traitement des langues, en m'appuyant sur l'expérience en cours d'espace libre de gestion de ressources linguistiques pour le français FReeBank.

3. Communications en lien avec cette table-ronde

Plusieurs communications présentées à ces JADT rejoignent la problématique abordée dans cette table-ronde, tout particulièrement :

Bécue M., Pagès J. et Pardo C.-E. Analysis of multilingual free responses.

Deroubaix J.-C. Que faire des corpus multilingues parallèles ? Une expérience.

Deville G., Dumortier L. et Paulussen H. Génération de corpus multilingues dans la mise en oeuvre d'un outil en ligne d'aide à la lecture de textes en langue étrangère.