Traitement automatisé du grec ancien et de l'arabe non classique à l'Institut orientaliste de l'Université catholique de Louvain

Bastien Kindt (kindt@ori.ucl.ac.be)
Laurence Tuerlinckx (tuerlinckx@ori.ucl.ac.be)

L'Institut orientaliste de l'Université catholique de Louvain (U.C.L., Louvain-la-Neuve, Belgique) partage ses activités entre l'enseignement et la recherche dans le domaine des langues, des littératures et des cultures orientales. L'Orient chrétien y est représenté, entre autres, par le *Centre d'Études sur Grégoire de Nazianze* dont l'ambition est de fournir l'édition critique des œuvres de cet auteur du IVe s. ap. J.-C., tant en grec que dans les différentes langues dans lesquelles ces textes ont été traduits, à savoir l'arménien, le syriaque, le géorgien et l'arabe (http://nazianzos.fltr.ucl.ac.be).

Dans ce contexte, les travaux faisant appel aux méthodes du TAL ont débuté avec la publication de la concordance lemmatisée des textes grecs de Grégoire de Nazianze. Les méthodes de travail et les outils développés ont été ensuite appliqués à d'autres auteurs, antérieurs ou postérieurs. Ces réalisations ont désormais pour cadre le *Projet de recherche en lexicologie grecque* (http://www.tpg.fltr.ucl.ac.be) dont la finalité est la constitution progressive d'un dictionnaire électronique par l'exploration systématique des sources patristiques et historiographiques byzantines. Un corpus de plus de 4.000.000 de mots-occurrences a été traité. Le dictionnaire compte plus de 700.000 « formes de mot » classées sous plus de 30.000 lemmes. Tous les développements (dictionnaire formes-lemmes, dictionnaire dérivationnel, concordanceur, lemmatiseur, graphes de levée automatisée de l'ambiguïté lexicale) sont conçus en collaboration avec le CENTAL (http://www.cental.fltr.ucl.ac.be). Dérivées d'UNITEX, un programme « open source » d'analyse lexicale et syntaxique des textes (http://www-igm.univ-mlv.fr/%7Eunitex), ces applications supportent le standard d'encodage UNICODE et permettront la mise au point de ressources accessibles aux utilisateurs extérieurs via Internet. Ces réalisations contribuent désormais à décrire de manière exhaustive et formalisée la langue grecque ancienne dans son ensemble. La première partie de l'exposé présentera et illustrera les capacités et limites de ces outils en matière d'analyse lexicale.

Les éditeurs de la version arabe de Grégoire de Nazianze ont souhaité développer un programme de lemmatisation de même type que celui du grec, dans le but d'accompagner les éditions d'index lemmatisés (et ultérieurement d'index bilingues grec-arabe). Cette version a été réalisée au Xe s. ap. J.-C. dans le milieu arabe chrétien d'Antioche, dans une langue qui se distingue de l'arabe classique par un certain nombre de particularités lexicales, morphologiques et syntaxiques relevant du « moyen arabe » (état de langue intermédiaire entre l'arabe classique et l'arabe moderne). Les éditions publiées par les membres du Centre respectent cet état de langue et fournissent un texte contenant des caractéristiques orthographiques et lexicales que les outils existants, conçus pour l'arabe standard moderne, ne peuvent efficacement prendre en compte. Une interface de lemmatisation a donc été mise au point (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_105.pdf). Par rapport au grec, l'arabe présente un grand nombre de formes graphiques correspondant à la concaténation de plusieurs unités lexicales distinctes. Le système développé prend cette réalité en charge, en proposant un lemme pour chacun des éléments constitutifs de la forme concaténée. Les formes non classiques (graphies non classiques, emprunts, termes proprement chrétiens ignorés dans les descriptions de l'arabe classique) sont intégrées dans le dictionnaire. Un corpus d'environ 21.000 « formes graphiques » est déjà lemmatisé. Les données lexicales du dictionnaire rassemblent plus de 32.000 « formes séparées » classées sous plus de 3.500 lemmes accompagnés d'une indication de leur nature et de leur racine. Créés dans le but de réaliser des index, ces outils (concordanceur, lemmatiseur, dictionnaire) sont désormais utilisés aussi pour des recherches plus larges directement utiles au travail d'édition et à l'étude de la langue. Ces réalisations fourniront la matière de la deuxième partie de l'exposé.

Le programme de lemmatisation arabe a été développé de manière à rester compatible avec les développements propres au grec. À moyen terme, les versions orientales pourront ainsi rejoindre la plateforme – dérivée d'UNITEX – qui permet le traitement des textes en grec ancien. La communication s'achèvera par des manipulations d'extraits de textes dans les différentes langues étudiées dans le cadre plus général des recherches du Centre.

Bibliographie du projet de lexicologie grecque

(Bibliographie complète sur le site du projet)

B. KINDT, *Avancées dans le traitement automatique du grec ancien à l'U.C.L.. L'analyse des textes au service d'une description lexicale de la langue. Une description lexicale de la*

*langue au service de l'analyse des texte*s, dans *Lexicometrica*, numéro spécial "Autour de la lemmatisation" (dir. D. LABBÉ) (2003), p. 1-17.
(http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema1.htm)

R. GÉRARD et B. KINDT, *D'un dictionnaire de lemmatisation (D.A.G.) à un Dictionnaire Dérivationnel Grec (D.D.G.),* dans *Le poids des mots. 7es Journées internationales d'Analyse statistique des Données Textuelles*, 10-12 mars 2004, Louvain-la-Neuve, ed. A. DISTER, C. FAIRON, G. PURNELLE, Louvain-la-Neuve, 2004, vol. I, p. 488-495.
(http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_046.pdf)

L. KEVERS et B. KINDT, Vers *un concordanceur-lemmatiseur en ligne du grec ancien*, dans *L'Antiquité Classique*, 73 (2004), p. 203-213.

B. KINDT, *La lemmatisation des sources patristiques et byzantines au service d'une description lexicale du grec ancien. Les principes de formulation des lemmes du Dictionnaire Automatique Grec*, dans *Byzantion*, 74 (2004), p. 213-272.

Bibliographie du projet de lemmatisation arabe

L. TUERLINCKX, *La lemmatisation de l'arabe non classique*, dans *Le poids des mots. Actes des 7èmes Journées Internationales d'Analyse statistique des Données Textuelles (Louvain-la-Neuve, 10-12 mars 2004)*, G. PURNELLE, C. FAIRON et A. DISTER eds, Louvain-la-Neuve, II, p. 1069-1078.
(http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_105.pdf)

Sites Web en rapport avec ces projets :

Site de l'Institut orientaliste de l'U.C.L., Louvain-la-Neuve :
http://www.fltr.ucl.ac.be/FLTR/GLOR/ORI/

Nazianzos : Site du Centre d'étude sur Grégoire de Nazianze à l'Institut orientaliste :
http://nazianzos.fltr.ucl.ac.be/

Site du projet de lexicologie grecque à l'Institut orientaliste :
http://www.fltr.ucl.ac.be/FLTR/GLOR/lexico/

Site du CENTAL, Centre de Traitement Automatique des Langues (U.C.L.) :
http://cental.fltr.ucl.ac.be/

Site d'UNITEX, un programme « open source » d'analyse lexicale et syntaxique des textes
http://www-igm.univ-mlv.fr/%7Eunitex

Site de Brepols Publishers (éditeur du Thesaurus Patrum Graecorum) :
http://www.brepols.net/