Laurent KEVERS
UCL – Cental
kevers@tedm.ucl.ac.be

Bastien KINDT
UCL – Institut orientaliste
kindt@ori.ucl.ac.be

Adaptation des ressources d'Unitex au traitement du grec ancien

Le traitement d'une nouvelle langue sous Unitex impose aux linguistes de conformer les ressources de ce programme aux particularités de la langue étudiée. Dans le cas du grec ancien, les difficultés suivantes sont apparues :
- les textes traités mêlent aux signes graphiques habituels des *signes critiques* utiles pour indiquer les problèmes d'établissement du texte mais ne relevant en aucun cas de la langue (par ex. crochets, angles, traits verticaux) ;
- pour une même forme de mot, les textes traités fournissent différentes graphies correspondant à des *conventions typographiques* variables selon les milieux ou les époques mais sans rapport avec l'évolution de la langue elle-même (par ex. iota souscrit ou adscrit) ;
- le grec ancien présente, pour une même forme de mot, un système complexe de *variations accentuelles* (changement d'accent, ajout d'accent) ;
- le grec ancien connaît les phénomènes d'*élision des formes simples*, tant à l'initiale qu'en finale ;
- le grec ancien connaît les phénomènes de *contraction* des mots (crases) ;
- etc.

L'exposé illustrera la répercussion de ces faits sur une analyse lexicale automatisée. Il indiquera ensuite les ressources du programme qui permettent d'y apporter des réponses efficaces et économiques. Ces dernières se situent au niveau des fichiers d'alphabet (Alphabet_sort.txt, Alphabet.txt), au niveau du dictionnaire (Delaf.dic), ou au niveau des graphes de prétraitement (Sentence.fst2, Replace.fst2). L'enjeu consiste donc à déterminer l'outil pertinent, à l'adapter et à le mettre en œuvre au moment opportun durant le processus d'analyse. L'exploration d'un texte adapté par les auteurs pour rassembler, en quelques lignes, un échantillon représentatif des difficultés décrites permettra de départager ce qui est acquis de ce qui ne l'est pas.

L'objectif poursuivi est l'étude du vocabulaire du grec ancien par le biais de la lemmatisation systématique des sources patristiques et historiographiques byzantines. Pour répondre à cette finalité précise, deux modules nouveaux ont été conçus et intégrés à Unitex : un module de référenciation et un module de lemmatisation. Le premier permet la récupération et l'affichage des références des mots dans le texte traité, chaque forme d'une concordance étant accompagnée de ses références complètes dans le texte. Le second permet d'attacher un lemme aux formes absentes du dictionnaire utilisé (le *Dictionnaire Automatique Grec*, outil propre au projet rassemblant les matériaux lexicaux acquis au fil des lemmatisations successives), ou de déterminer le lemme pertinent in textu pour une forme ambiguë dans le dictionnaire. Ces deux modules feront également l'objet d'une présentation.

Au-delà des applications directement attendues par les concepteurs du projet, ces travaux ouvrent deux perspectives plus générales :
- une étude contrastive du grec ancien et du grec moderne, puisque des réalités linguistiques communes aux deux états de cette langue pourront désormais être traitées sous le même environnement ;

- la création d'outils périphériques à Unitex susceptibles d'être utiles aussi pour l'analyse d'autres langues.

Références

Kevers, Laurent ; Bastien Kindt. 2004. Vers un concordanceur-lemmatiseur en ligne du grec ancien. In *L'Antiquité Classique* 73, pp. 203-213.

Kevers, Laurent ; Bastien Kindt. 2005. Traitement automatisé de l'ambiguïté lexicale en grec ancien. Première approche par application de grammaires locales. In *Lingvisticae Investigationes* [à paraître].

Kindt, Bastien. 2004. 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 (*D.A.G.*). In *Byzantion* 74, pp. 213-272.