Traitement automatisé de l'ambiguïté lexicale en grec ancien. Première approche par application de grammaires locales

Laurent Kevers (UCL - CENTAL)
Bastien Kindt (UCL - Institut orientaliste)

Le Dictionnaire Automatique Grec (DAG) est un dictionnaire électronique du grec ancien. Une de ses versions actuelles, répondant au standard UNICODE et mise au format DELAF, totalise 221.234 entrées-formes accrues de leur lemme et de codes d'analyse, limités pour l'instant à la catégorie morphosyntaxique. Utilisé pour fournir un étiquetage lexical des corpus sur lesquels il est appliqué, le DAG intervient dans le processus d'élaboration automatisée de concordances lemmatisées de sources grecques patristiques ou historiques d'époque byzantine.

Une production récente portait sur l'analyse des œuvres complètes d'un auteur du IVe s. av. J.-C., Basile de Césarée (707.853 occurrences, 14.843 lemmes et 68.867 formes de mots). À cette occasion, de dictionnaire a couvert 93,85% du vocabulaire du corpus. La proportion des formes homographes ayant reçu plus d'une proposition de lemme s'élevait cependant à 11,78% du nombre total d'occurrences. Une intervention manuelle fastidieuse et coûteuse, mais indispensable pour offrir aux utilisateurs des concordances des données entièrement désambiguïsées, a donc fait suite à la phase automatisée du traitement.

L'objet de cette contribution est de présenter une première série de grammaires locales susceptibles de faciliter l'achèvement de la lemmatisation par un traitement assisté par ordinateur des ambiguïtés. La première partie dresse un tableau qualitatif et quantitatif des ambiguïtés lexicales du grec ancien, illustré sur base des cas rencontrés dans le corpus et dans le dictionnaire. La deuxième décrit les grammaires utilisées, les postulats sur lesquels elles reposent, leur mode d'application au texte. Elle analyse ensuite les résultats obtenus après leur application au texte, non lemmatisé, de la correspondance de Basile de Césarée (une partie donc du corpus entier de l'auteur; 134.511 occurrences, 7.488 lemmes et 24.815 formes de mots). La dernière partie dresse le bilan des ces premiers travaux: les résultats obtenus par application des grammaires sont confrontés aux données du texte désambiguïsé jadis manuellement et les orientations futurs du projet sont envisagées.