Visualisation des motifs séquentiels extraits à partir d'un corpus en Ancien Français

Julien Rabatel*, Yuan Lin*, Yoann Pitarch*, Hassan Saneifar*,
Claire Serp**, Mathieu Roche*, Anne Laurent*

*LIRMM, Université Montpellier 2 - CNRS UMR5506,
{mroche,laurent}@lirmm.fr
**Université Montpellier 3,
serpclaire@yahoo.fr

Résumé. Cet article présente une interface permettant de visualiser des motifs séquentiels extraits à partir de données textuelles en Ancien Français.

1 Introduction

Les travaux présentés dans cet article répondent aux besoins d'une experte médiéviste souhaitant découvrir des connaissances nouvelles dans un corpus de textes écrits en Ancien Français. Les connaissances extraites à partir de ce corpus sont sous forme de motifs séquentiels. Dans notre contexte, un motif séquentiel est une suite ordonnée d'itemsets (phrases). Un itemset est un ensemble d'items (mots). Par exemple, le motif <(chevalier dam)(roi)> extrait à partir de notre corpus signifie que, souvent, les mots "chevalier" et "dam" apparaissent ensemble au sein d'une même phrase avant l'apparition de "roi" dans une phrase suivante. Ceci permet aux experts d'analyser, sans *a priori*, les mots et enchaînements de mots qui apparaissent dans un même contexte, mettant ainsi en relief des associations susceptibles d'apporter des connaissances nouvelles à un expert. Notons que dans l'étude actuellement menée, l'experte médiéviste souhaite plus particulièrement découvrir des motifs séquentiels faisant intervenir des mots propres à la parenté. Les différentes étapes et fonctionnalités de notre logiciel sont décrites dans la section suivante.

2 Processus d'extraction des motifs séquentiels

La première étape du prétraitement des données textuelles consiste à appliquer le Tree Tagger de Schmid (1994) qui possède des règles et des lexiques adaptés à l'Ancien Français. Ce système apporte des informations grammaticales aux différents mots du texte (par exemple, étiquettes "adjectif", "nom", etc). Les mots qui sont davantage porteurs de sens tels que les noms peuvent alors être filtrés. Par ailleurs, l'utilisation du Tree Tagger permet de lemmatiser les mots du corpus. Après ce prétraitement, l'extraction des motifs séquentiels à partir des données textuelles peut s'effectuer à l'aide de la méthode SPaC (Sequential PAtterns for Text Classification) qui est décrite dans (Jaillet et al. (2006)).

Un thème pouvant être privilégié par l'utilisateur (dans notre cas la parenté), notre logiciel permet de n'extraire que des motifs relatifs à cette thématique au travers d'une liste de