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

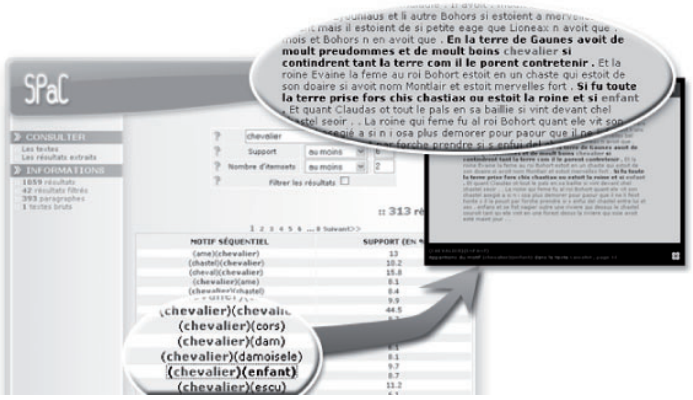

FIG. 1 – *Le moteur de recherche de l'interface de visualisation des résultats et la consultation de l'origine du motif séquentiel (chevalier)(enfant).*

mots pertinents du domaine. Les motifs séquentiels extraits du corpus ne seront alors que ceux dont au moins un item est un mot de la liste manuellement établie pas l'utilisateur. L'ajout de connaissances permet donc de filtrer les motifs offrant ainsi à l'utilisateur des informations à la fois complètes (recherche sur l'ensemble du corpus) et pertinentes (adaptées à la thématique).

Afin de répondre à la difficulté liée au nombre de motifs qui peut être élevé, notre application s'accompagne d'un moteur de recherche permettant de mettre en relief les motifs contenant un ou plusieurs mots spécifiés par l'utilisateur (figure 1). Les résultats d'une recherche peuvent également être triés selon plusieurs critères (support, nombre d'itemsets). Par ailleurs, notre logiciel permet de visualiser les phrases qui valident un motif séquentiel donné.

3 Conclusion

Nous proposons dans cet article un ensemble de méthodes implantées au sein d'une interface dédiée aux utilisateurs non informaticiens mais experts des données. Une perspective envisageable pourrait exploiter le fait que l'approche présentée s'avère adaptée à une classification des paragraphes du corpus en fonction des thématiques présentes dans ce dernier.

Références

Jaillet, S., A. Laurent, et M. Teisseire (2006). Sequential Patterns for Text Categorization. *International Journal of Intelligent Data Analysis (IDA) 10*(3).

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49.

Summary

This paper introduces a tool to visualize sequential patterns extracted from textual data in Old French.