La quête du Graal et la réalité numérique

Claire Serp, Anne Laurent, Mathieu Roche, Maguelonne Teisseire

La quête du Graal et la réalité numérique

Claire SERP
Université Montpellier 3
serpclaire@yahoo.fr

Anne LAURENT, Mathieu ROCHE,
Maguelonne TEISSEIRE
LIRMM, Université Montpellier 2 - CNRS UMR5506
{laurent, mroche, teisseire}@lirmm.fr

Résumé.

Cet article présente un processus de fouille de données afin d'extraire des connaissances associées au motif de la parenté et de la famille dans un corpus en ancien français de la première moitié du XIIIe siècle. Après une numérisation puis un prétraitement des données fondé sur des techniques de TAL (Traitement Automatique du Langage), il a été procédé à une extraction de motifs séquentiels (enchaînements de voisinages de mots liés à la thématique traitée). Dans cet article, nous présentons les problèmes liés à la numérisation et aux traitements du texte. Puis, nous détaillons ce processus automatique et exhaustif et analysons les premiers résultats obtenus en procédant à l'étude précise d'un motif séquentiel dans différents textes du cycle *Lancelot-Graal*[1].

[1] Alexandre Micha, *Lancelot, roman en prose du XIIIe siècle*, Genève, Librairie Droz, 9 volumes, 1978 à 1982, *La Queste del Saint Graal*, publiée par Albert Pauphilet, Honoré Champion, Paris, 2003 (identifiant BFM : qgraal) *La Mort le roi Artu*, Jean Frappier, Droz, Genève, 1996 (identifiant BFM : artu).

Mots clés. Fouille de Données, Motifs Séquentiels, Traitement Automatique des Langues (TAL), Etiquetage Grammatical, Ancien Français.

Abstract.

This paper describes a method to mine texts written in Old French in the second half of the 13th century. This method is based on data mining tools, which are used here to automatically extract patterns dealing with family relationships. After scanning and digitizing the texts, a pretreatment has been applied based on NLP (Natural Language Processing) to clean the texts. Sequential patterns are then extracted, which provide the expert with sequences of neighborhoods of words that are relevant to the analysis being carried out by the expert. We report here the problems raised by numerizing such texts. Then we present the process we have applied and the results we have obtained. Finally, we detail how one sequential pattern, chosen among those appearing in the *Lancelot-Graal series*, can be further analyzed.

Keywords. Data Mining, Sequential Patterns, Natural Language Processing (NLP), Part-of-speech Tagging, Old French.

1- Contexte et problématique

Les travaux présentés dans cet article s'intéressent à la découverte de connaissances nouvelles dans un corpus écrit en français médiéval, le cycle *Lancelot-Graal*, un vaste ensemble de textes écrits dans la première moitié du XIIIe siècle, comprenant cinq ouvrages, le *Joseph d'Arimathie*, le *Merlin* de Robert de Boron, le *Lancelot en prose*, *La Queste del Saint Graal* et *La Mort le roi Artu*. Outre une évidente homogénéité thématique, F. Lot relève que « D'un bout à l'autre de l'Estoire à la Mort d'Arthur, langue et style sont identiques. D'un bout à

l'autre on rencontre les mêmes particularités, les mêmes manies d'écrivain (…)[2] [3]» Dans le cadre d'un travail de doctorat sur le thème « Identité, filiation et problèmes de parenté dans les romans du Graal en prose », la question de la numérisation de ces textes s'est posée.

Bien qu'étant une source de données extrêmement riche, il existe peu de travaux automatisés sur le traitement des textes en ancien français, principalement du fait de la rareté des ressources numériques. Dans ce domaine, il est important de citer les travaux de la BFM[4] (Base de Français Médiéval) diffusés par le système Weblex[5] développé au laboratoire ICAR de l'Ecole Normale Supérieure de Lettres et Sciences Humaines (Heiden, 2004), (Heiden et Lavrentiev, 2004). Les textes ont été normalisés selon le format XML[6] permettant d'obtenir le document selon une structure arborescente avec les méta-données associées. La recherche d'information, à l'aide d'un moteur de recherche, offre alors la possibilité de faire des recherches de mots ou de successions de mots. L'utilisateur peut procéder à des recherches sur les textes, en fonction de spécification endogène ou exogène, et peut croiser un certain nombre d'informations.

Notre démarche diffère de ces travaux car elle se situe dans le contexte de l'extraction de connaissances sans a priori. En effet, à partir de la numérisation d'un roman médiéval nous abordons ses différentes exploitations, notamment l'application d'un processus de fouille de données afin d'en extraire des connaissances. Ainsi, après une numérisation, nous appliquons un prétraitement des données fondé sur des techniques de TAL

---

2 Au sujet de la cohérence du cycle et sa constitution en tant que corpus, voir Ferdinand Lot, *Etude sur le Lancelot en Prose*, Honoré Champion, Paris, 1984, chap. IV « Unité de plan et d'esprit », en particulier partie D, sur la langue et le style, p. 65 à 107.

3 Ibid, p. 107.

4 http://w3.ens-lsh.fr/egerstenkorn/bfm2/

5 http://weblex.ens-lsh.fr/wlx/

6 Pour la justification de ce format, voir l'article de (Heiden et Guillot, 2003)

(Traitement Automatique du Langage) dont l'objectif final est de procéder à une extraction de motifs séquentiels (enchaînements de voisinages de mots au fil du document). Ceux-ci sont ensuite analysés et validés par l'experte en littérature médiévale ce qui permet alors d'obtenir une connaissance qui n'aurait pas pu être obtenue selon une approche dirigée.

Cet article détaille notre proposition et est organisé de la façon suivante : nous décrivons tout d'abord la phase d'acquisition du corpus pour détailler ensuite le processus de prétraitement et l'étape de fouille de données. Nous présentons alors les résultats obtenus en soulignant la démarche non dirigée de cette approche qui conduit en particulier à des découvertes surprenantes. Nous concluons par un bilan et les nombreuses perspectives offertes par ces premières analyses.

2- Acquisition du corpus

La numérisation du texte présente un triple intérêt. Tout d'abord, elle permet une circulation rapide dans le texte, avec des recherches automatiques, ensuite elle améliore le relevé et la gestion des occurrences (mots, syntagmes, etc), enfin, elle permet de tenter d'appliquer des outils d'analyse innovants.

La numérisation fait l'objet d'études spécifiques. Citons en particulier le CCFM[7] (Consortium international pour les corpus de français médiéval), créé en 2004, qui vise à proposer un processus rigoureux et unifié pour l'acquisition de corpus médiévaux.

Le thème de recherche « Identité, filiation et problèmes de parenté dans les romans du Graal en prose » et la numérisation du corpus associé impliquaient deux difficultés majeures : d'une part l'ampleur du corpus (le *Lancelot en prose* qui est composé de 8 tomes, représentant plus de 750 000 mots)

[7] http://ccfm.ens-lsh.fr/spip.php?rubrique8

et d'autre part le nombre important des entrées lexicales constituant les points de départ des analyses à mener. Le *Lancelot en prose* est un vaste roman, écrit dans la première moitié du XIIIe siècle.

Cette numérisation n'est pas sans poser un certain nombre de problèmes. Pour pouvoir procéder à l'extraction de connaissances de façon automatique, il est nécessaire de disposer d'un texte normalisé. Par exemple, lors de la transcription du texte papier au texte numérique, les notes de bas de pages de l'édition d'A. Micha se retrouvent en corps de texte. Elles sont souvent de très mauvaise qualité et les appels de notes sont mal intégrés dans le texte. La solution serait une suppression pure et simple à la fois des appels de notes mais aussi des notes elles-mêmes. Mais pour le chercheur en littérature, ces indications sont essentielles, puisqu'elles répertorient les variantes de manuscrits, les rétablissements de textes (et donc d'une certaine façon les modifications) effectués par le critique qui a fait cette édition. La solution retenue est donc deux versions du même texte, d'une part une version gardant les notes, même de mauvaise qualité, d'autre part une version numérique non bruitée, pour pouvoir effectuer un traitement pertinent.

3- Le processus d'extraction des connaissances

Avant d'appliquer un processus de fouille de données sur les textes, un prétraitement qui s'appuie sur des outils de TAL peut être mis en place. Pour ces différents prétraitements, nous avons utilisé le Tree Tagger (Schmid, 1994) car cet outil s'appuie sur des ressources adaptées à l'ancien français. L'extraction des connaissances sous forme de motifs séquentiels à partir des textes prétraités s'effectue grâce à la méthode SPaC (Jaillet *et al.*, 2006). Une interface graphique Java a été ajoutée pour permettre à l'expert d'exécuter facilement le processus d'extraction (Rabatel *et al.*, 2008).

3.1 Prétraitement des données par étiquetage grammatical

Après l'étape d'acquisition du corpus, la phase suivante du processus de fouille de textes consiste à étiqueter les documents numérisés. Ce traitement permet d'apposer une étiquette morpho-syntaxique à chacun des mots du texte (noms, verbes, etc). Plusieurs étiqueteurs plus ou moins adaptés à l'étude des textes médiévaux peuvent être utilisés. Nous citerons tout d'abord l'étiqueteur de Brill (Brill, 1994) qui s'appuie sur des lexiques et des règles lexicales et contextuelles pour déterminer des étiquettes adaptées. Un tel étiqueteur est utilisé dans le cadre des travaux de la BFM (Prévost et Heiden, 2004). Nous pouvons également citer l'étiqueteur Tree Tagger conçu par H. Schmid (Schmid, 1994). Cet étiqueteur que nous allons utiliser dans nos travaux s'appuie sur des résultats probabilistes afin de prédire les étiquettes morpho-syntaxiques à appliquer. Les probabilités sont « apprises » récursivement à partir d'un ensemble de trigrammes connus (suites de trois étiquettes grammaticales consécutives constituant l'ensemble d'apprentissage). Une méthode d'apprentissage automatique consistant à déterminer une fonction de prédiction à partir des données expertisées est alors mise en œuvre. À chaque étape d'apprentissage, le résultat de la fonction de prédiction est évalué. Tant qu'un nombre d'erreurs seuil subsiste, le processus d'apprentissage continue. Le lecteur intéressé pourra se référer à l'ouvrage Apprentissage Artificiel (Cornuéjols et Miclet, 2002) qui décrit de manière approfondie différents algorithmes d'apprentissage tels que les arbres de décision binaires qui sont utilisés par le Tree Tagger. Notons que d'autres approches consistant à l'analyse des données textuelles en ancien français s'appuient uniquement sur des approches statistiques sans catégorisation morpho-syntaxique préalable (Dupuis et Lebart, 2008) contrairement à nos travaux.

Lorsque le processus d'apprentissage est terminé, la fonction de prédiction apprise peut être appliquée à de nouveaux textes. Les expérimentations menées dont les résultats ont été présentés dans (Serp *et al.*, 2008) ont montré un taux de réussite (précision) de l'ordre de 80% à partir du corpus que nous

étudions. Notons que cette précision reste inférieure aux valeurs obtenues dans les travaux de (Stein, 2003) (de l'ordre de 95%). Ceci peut s'expliquer par les caractéristiques différentes (lexicales, syntaxiques, etc) de notre corpus comparativement aux données décrites ci-dessous qui sont utilisées par le Tree Tagger pour apprendre les règles d'étiquetage.

Afin d'apposer les étiquettes morpho-syntaxiques, le Tree Tagger s'appuie sur des connaissances lexicales de l'ancien français (Stein, 2003). Le lexique utilisé par le Tree Tagger provient d'un ensemble de ressources dont la plus significative est le *Corpus d'Amsterdam*. Ce corpus a été établi au début des années 80 par un groupe de chercheurs sous la direction d'Anthonij Dees et il est à la base de la publication de *l'Atlas des formes linguistiques des textes littéraires de l'ancien français*. Ainsi, les règles d'étiquetage et un lexique adapté à l'ancien français ont pu être constitués. Ce *Corpus d'Amsterdam* regroupe 289 textes différents ce qui a permis d'obtenir un premier lexique de plus de 130 000 élements. À cela s'ajoute diverses ressources lexicales telles que la version électronique de *l'Altfranzösisches Wörterbuch d'A. Tobler et E. Lommatzsch* ou encore *les fiches linguistiques de M. Robert Martin* (ATILF). Au final, le lexique en ancien français que nous avons utilisé contient un peu plus de 234 000 éléments.

La figure ci-dessous illustre le résultat de l'étiquetage de la phrase suivante qui est issue de notre corpus : « En la marche de Gaule et de la petite Bertaigne avoit ».

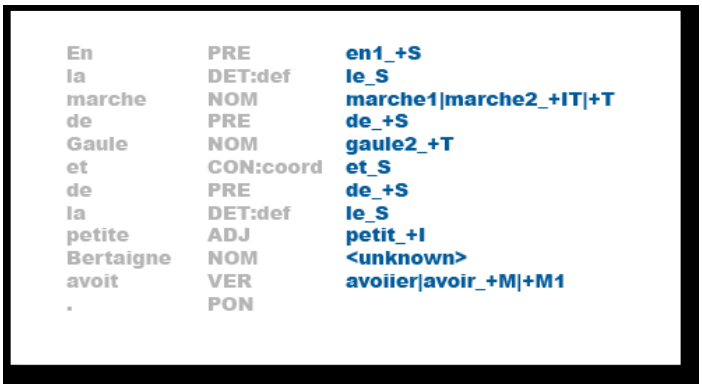

Figure 1 – Résultat de la phase d'étiquetage

La première colonne correspond aux termes de la phrase, la seconde nous renseigne sur la catégorie grammaticale de ces termes et la dernière nous donne leur forme lemmatisée, aussi appelée forme canonique (verbes à l'infinitif, noms au masculin singulier, etc).

En français médiéval de nombreux lemmes possèdent des variantes graphiques. Le Tree Tagger prend en compte différentes formes comme le mot « avoit » qui a deux possibilités de lemmes (avoiier, avoir) illustré sur la figure 1. La sortie du Tree Tagger donne l'origine des formes de lemmes par la présence de caractères en majuscule (dernière colonne) : "_M" pour les formes de R. Martin, "_T" pour Tobler-Lommatzsch, etc. Le détail de ce format est donné dans (Stein, 2003). Par ailleurs, l'application que nous avons développée qui a été présentée dans (Rabatel *et al.*, 2008) permet d'utiliser d'autres variantes graphiques déterminées manuellement (nous pouvons par exemple citer le mot « sœur » que l'on peut trouver dans le tome VII du *Lancelot* sous les formes suivantes : *soeur, serours, seur, seror, seurs, serours, suer*).

L'utilisation de l'étiquetage permet d'effectuer deux types de prétraitements essentiels pour la phase suivante du processus de fouille de données :

(1) lemmatiser les textes pour rassembler des mots et ainsi améliorer le processus d'extraction de connaissances sous forme de motifs séquentiels qui est détaillé dans la section suivante.

(2) filtrer les mots ayant une étiquette morpho-syntaxique précise présents dans les motifs. Ceci permet par exemple d'extraire seulement les motifs composés de noms souvent beaucoup plus porteurs de sens. Les mots de type adjectif seront également filtrés permettant une analyse complète décrite dans la section 4 de cet article.

Ces deux prétraitements permettent d'extraire des informations plus pertinentes et plus précises dans le cadre de la thématique de la parenté qui est en particulier développée dans le cadre de ce projet.

3.2 Les motifs séquentiels

Les motifs séquentiels sont l'une des méthodes de fouille de données les plus utilisées au sein du processus d'extraction de connaissances à partir de données (ECD) qui consiste à extraire des connaissances pertinentes et nouvelles à partir de gros volumes de données. Ce processus enchaîne différents traitements dont principalement : la sélection des données à traiter au sein de sources de données hétérogènes (documents textes, documents excel, documents internet, bases de données, etc), le nettoyage de ces données (e.g. homogénéisation des unités), la fouille de données elle-même, et enfin l'interprétation des résultats par un expert (littéraire par exemple dans notre cas).

L'extraction de motifs séquentiels permet de découvrir des connaissances sous la forme de règles qui décrivent les grandes tendances présentes dans les bases de données. Un motif séquentiel est par exemple « 28% de clients de cette librairie achètent d'abord le roman X, puis le roman Y accompagné de la bande dessinée A, puis enfin le livre illustré Z ». Cet exemple met en avant plusieurs caractéristiques de ce type de règle. Nous relèverons notamment la présence d'un terme indiquant à quel point cette règle est présente dans les données, ici « 28% ». On parle de support, souvent exprimé par un pourcentage comme c'est le cas ici. Ce support est lié au nombre d'objets dans la base vérifiant la règle. Les objets (ici les clients de la librairie) sont donc au cœur de l'extraction. Cependant, les données apparaissant dans les règles sont autres. Il s'agit des attributs de la base de données que l'on cherche à étudier. On parle d'items quand un seul attribut est présent (par exemple « roman X » ou « bande dessinée A ») et d'itemset quand plusieurs attributs sont présents en même temps (« roman X et bande dessinée A »). Enfin, il est crucial de noter la

présence d'un ordre d'apparition (ici chronologique) de ces itemsets.

Pour extraire de tels motifs séquentiels, il faut donc être muni de bases de données contenant l'ensemble des informations décrites ci-dessus :

- des clients (que l'on comptera),
- des items (que l'on fera apparaître dans les motifs extraits, sous la forme d'itemsets quand plusieurs items seront concernés),
- des dates d'apparition de ces items (qui serviront à ordonner les itemsets dans les motifs).

Enfin, il faut avoir défini le seuil minimal de support à partir duquel il est intéressant de considérer les motifs, qui sera souvent noté *minsup*. On parle de motif fréquent quand le support est supérieur à ce seuil *minsup*. Le problème de l'extraction de motifs séquentiels revient alors à extraire tous les motifs les plus longs possibles ayant un support supérieur au seuil *minsup*. Ce seuil permet non seulement de ne considérer des motifs que s'ils sont représentatifs au niveau de la base de données, mais permet également de construire des algorithmes performants. En effet, il est facile de constater que nous sommes en présence d'une propriété anti-monotone : quel que soit le motif, plus on lui ajoute des items, plus son support diminue. Par exemple, si 25% des clients ont acheté le roman X puis le roman Y, alors la proportion de clients ayant acheté le roman X puis le roman Y et la bande dessinée A est forcément inférieur à 25%.

Notons que l'ajout d'un item à un motif peut se faire selon deux manières :

- soit au sein d'un itemset existant (c'est-à-dire à une date déjà présente dans le motif), comme nous l'avons fait ci-dessus en ajoutant « et la bande dessinée A »,

- soit en ajoutant un itemset en fin de motif, comme nous aurions pu le faire en considérant le motif « roman X puis roman Y puis la bande dessinée A ».

Cette propriété permet donc de construire des motifs grâce à un algorithme performant qui évite tous les chemins de construction de motifs qui ne peuvent pas être fréquents. Le lecteur intéressé pourra se référer aux articles de la littérature pour plus d'information (Masseglia *et al.*, 2004).

Les motifs séquentiels ont été utilisés dans de très nombreuses applications, notamment dans le contexte de l'étude du panier de la ménagère (tendances des produits vendus), du web usage mining (étude des chemins de navigation des internautes, notamment à des fins marketing ou d'amélioration de l'ergonomie). Ils ont également été appliqués aux données textuelles (Jaillet *et al.*, 2006). Par exemple, la méthode SPaC permet l'extraction de motifs séquentiels décrivant les textes, en considérant :

- les objets (clients) comme étant les textes ou les parties de textes,
- les items comme étant les mots d'un texte,
- les itemsets comme étant les ensembles de mots présents ensemble au sein d'une phrase ou d'un paragraphe,
- les dates comme étant l'ordre d'apparition d'une phrase au sein d'un texte.

Ceci permet d'extraire des motifs du type « 38% des textes contiennent les mots 'élégant' et 'frère' au sein d'une phrase puis le mot 'charisme' dans une phrase suivante ». Ce motif est noté <(élégant, frère)(charisme)>.

Notons que dans les motifs séquentiels classiques, la distance entre un itemset et le suivant n'est pas contrainte. Il n'y a donc ici aucune information sur le nombre de phrases qui séparent la phrase dans laquelle les mots 'élégant' et 'frère'

apparaissent et la phrase dans laquelle ‘charisme’ apparaît. Les motifs avec contraintes d’espacement entre itemsets ont donc été introduits et permettent à l’utilisateur d’imposer des bornes minimales et/ou maximales (Masseglia *et al.*, 2004).

Notons que l'approche présentée dans cet article considère chaque itemset comme un *sac de mots* sans véritablement considérer les syntagmes pertinents (Jacquemin et Bourigault, 2003). La prise en compte des syntagmes nominaux est un premier traitement que nous avons étudié dans (Serp *et al.*, 2008). Contrairement aux approches de lexicométrie tendant à retrouver des voisinages (Mayaffre, 2007), notre approche vise à mettre en relief des *enchaînements de voisinages* fréquemment retrouvés au fil du corpus. Ces enchaînements peuvent mettre en valeur des connaissances nouvelles en considérant différents types d’itemsets (phrases, paragraphes, etc).

4. Analyse des résultats

Dans nos travaux consistant à appliquer un processus de fouille de données à un texte médiéval, notre premier objectif est de vérifier que ce processus est adapté et que les motifs séquentiels extraits sont pertinents. Précisons que les motifs ont été extraits après avoir appliqué les traitements linguistiques détaillés dans la partie 3 (lemmatisation des mots du corpus et filtrage des noms).

Une fois ces motifs extraits il est alors pertinent de se focaliser sur un ou plusieurs motifs comme élément d’analyse linguistique. L’extraction permet alors de valider des hypothèses par la mise en relation de différents motifs ou encore de faire émerger de nouvelles représentations.

L’interface d’extraction de motifs séquentiels développée dans le cadre de ce projet est illustrée sur la figure suivante :

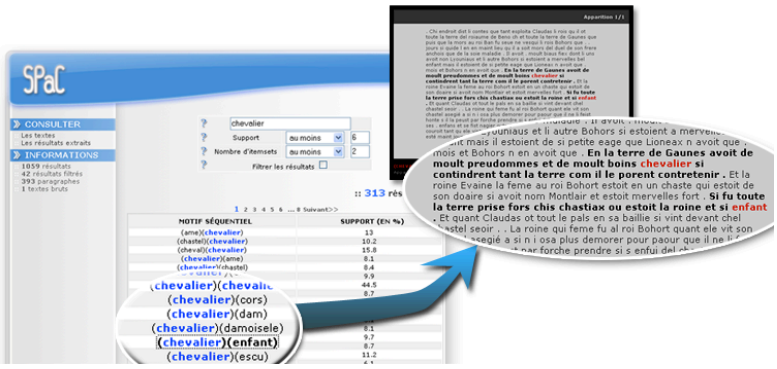

Figure 2 – L'interface SPaC

Lors de l'extraction de motifs séquentiels, le mot « chose » est apparu comme étant un motif séquentiel important, dans la mesure où il apparaît avec un support de 47,6% dans le *Lancelot*, 46,9% dans *La Queste Del Saint Grall*, et de 36,8% dans La Mort le roi Artu.

Le mot « chose » est un mot complexe qui peut avoir un grand nombre de sens en fonction du contexte. Cette analyse a été réalisée en comparant les motifs séquentiels de deux textes du cycle, le *Lancelot*[8] et *La Queste del Saint Graal*[9].

Dans le *Lancelot*, les occurrences du mot « chose » servent généralement à assurer une continuation thématique, en fonctionnant comme référents anaphoriques[10]. On le trouve souvent associé au verbe « paroler » ou « dire[11] ». Il est très souvent associé au mot « chevalier » : <(chevalier chose)> 7,8% et <(chevalier)(chose)> 13,9%. Le pourcentage d'apparition n'est pas le même car dans ce dernier motif, les mots « chevalier » et « chose » ne sont pas retrouvés au sein de la même phrase.

---

8 Sur un échantillon du texte correspondant à 390 pages.

9 La Queste del Saint Graal, publié par A. Pauphilet, Honoré Champion, Paris, 2003.

10 ex page 211 « Toues ces chose ot bien oïes la damoisele »

11 ex page 266 « chose k'il l'en die ».

Nous trouvons également <(chose)(chevalier)> 12,8%, <(chevalier chose)(chevalier)> 4,2%, mais aussi aux termes renvoyant à la femme et à l'amour :

<(amor)(chose)>[12] 7,2%, <(chose)(cuer)>[13] 10,6%,
<(chose)(dam)>[14] 25,4%, <(chose)(damoisele)>[15] 33,1%,
<(chose)(roine)>[16] 20,3%.

Ce motif est parfois associé au verbe *aimer* (« chose que le plus aim »). Le terme de *chose* a donc une valeur anaphorique, qui sert à résumer ce qui précède, ou alors il prend le sens d' « objet » que l'on aime ou que l'on défend.

Dans La Queste, contrairement au texte précédent, on retrouve une très faible représentation de l'association de la femme (dame et demoiselle) et du terme « chose », puisque ce motif n'est présent que sous la forme <(damoisele)(chose)> et n'a qu'un support de 1,5%.

Les supports d'une partie des motifs incluant le mot « chose » extraits à partir de La Queste del Saint Graal sont présentés dans le tableau 1.

MOTIF SEQUENTIEL	SUPPORT (EN %)
<(chose)>	46.9
<(chose dieu)>	3.4
<(chose foi)>	3.1

[12] Sous les formes : <(amor)(chose)> 3,6%, <(chose)(amor)> 3,6%.

[13] Sous les formes : <(chose)(cuer)> 5,6%, <(chose cuer)> 5%.

[14] Sous les formes : <(chose dam)> 5,6%, <(chose)(dam)> 9,5%, <(dam)(chose)> 10,3%.

[15] Sous les formes : <(chose damoisele)> 5,8%, <(chose)(damoisele)> 7,8%, <(damoisele)(chose)> 8,6%, <(chevalier damoisele)(chose)> 4,2%, <(chevalier)(chose)(damoisele)> 3,1%, <(chose damoisele)(damoisele)>3,6%.

[16] Sous les formes : <(chose roine)> 4,7%, <(chose)(roine)> 8,1%, <(roine)(chose)> 7,5%.

MOTIF SEQUENTIEL	SUPPORT (EN %)
<(chose parole)>	3.4
<(chose saint)>	3.4
<(chose seignor)>	5
<(chose verite)>	3.1
<(chose)(ame)>	5
<(chose)(arbre)>	3.8
<(chose)(aventure)>	6.9
<(chose)(avis)>	3.4
<(chose)(chaiste)>	5
<(chose)(cheval)>	5.7
<(chose)(chevalerie)>	3.1
<(chose)(dieu)>	7.6
<(chose)(espee)>	5.3

Tableau 1 – Quelques motifs et leur support

On retrouve également le mot « chose » pour construire une continuation thématique, mais dans une moindre proportion. Un emploi nouveau apparaît, mis à jour par l'étude des motifs séquentiels sur l'ensemble du corpus, puisqu'on le trouve associé à des termes religieux : (chose) et (saint)[17] 14,8% (chose) et (ciel)[18]3,1%, (chose) et (dieu)[19]17,5% (chose) et (frere)[20] 9,5% (chose) et (foi) 13,8% [21]

La « chose » en question n'est plus en rapport avec la femme. Alors quelle est cette chose ? Une étude approfondie

[17] Sous les formes <(chose saint)> 3,4%, <(chose)(saint)> 6,1%, <(saint)(chose)>5,3%.

[18] Sous les formes <(chose)(ciel)> 3,1%.

[19] Sous les formes <(chose dieu)> 3,4%, <(chose)(dieu)> 7,6%, <(dieu)(chose)> 6,5%. On pourrait également dans certains cas ajouter le terme de « seignor » qui renvoie parfois à Dieu ou au Christ.

[20] (Frère dans le sens de membre d'un ordre religieux). Sous les formes <(chose)(frere)> 5,7%, <(frere)(chose)> 3,8%.

[21] Sous les formes <(chose foi)> 3,1%, <(chose)(foi)> 3,8%, <(foi)(chose)> 6,9%.

des adjectifs associés fait apparaître les termes de « esperiteux[22] » et « celeste », termes souvent employés en opposition aux « terriennes choses » : « car la Queste n'est mie de terrianes choses, mes de celestielx[23]». C'est donc du côté de la quête du Saint Vase qu'il faut chercher l'explication de ces motifs séquentiels. Le mot « chose » renvoie donc souvent au mystère du Graal, comme le montre cette occurrence « ce est li Sainz Graax, les secrees choses Nostre Seignor[24] ». Ce qui explique alors l'association de (chose) et de (frère). Seule l'Eglise peut donner l'explication des mystères sacrés, et en donner la *senefiance*[25]. L'association du terme de (chose) et de celui de (senefiance) est d'ailleurs présente dans ce texte, alors qu'il ne l'est pas dans le premier[26]. Le frère est donc celui qui sert de médiation entre Dieu et ses soldats, et qui explique, dans la mesure du possible le sens secret des évènements. Cette association entre le mot « chose » et le vocabulaire religieux montre alors que ce mot n'est plus utilisé pour résumer quelque chose, mais bien pour pallier l'absence de signifiant. Car les mystères du Graal relèvent de l'indicible.

Cette étude d'un motif montre bien qu'un même terme en fonction des auteurs et de la perspective d'un texte, peut assumer différents sens et différentes fonctions. Mot « fourre-tout » dans le *Lancelot*, il devient dans *La Quête du Graal* le symbole même de l'impuissance des hommes à appréhender le divin.

Lorsque enfin le héros contemplera le Saint Vase, il ne pourra que s'exclamer « ore voi ge tot apertement ce que langage ne porroit descrire ne cuer penser[27] », laissant définitivement en suspend le mystère de cette étrange… chose.

---

22 Egalement sous la forme « esperitiex ».

23 p 127. Autre exemple p 129.

24 p 158.

25 Le sens, l'explication.

26 Sous la forme <(chose)(senefiance)> 2,5%.

27 La Queste p 278 « je vois désormais clairement ce que le langage ne peut décrire et l'esprit ne peut concevoir ».

5. Conclusion

Le processus de numérisation des textes médiévaux prend de plus en plus d'ampleur. Qu'il s'agisse de textes mis à la disposition des étudiants (numérisation par exemple des textes au programme de l'agrégation) ou de travaux effectués dans les centres de recherche. Dans ce contexte, nous avons mis en œuvre un processus d'extraction de connaissances sur un corpus médiéval de grande ampleur. Partant de la numérisation des documents, nous avons appliqué des prétraitements efficaces fondés sur des techniques propres au traitement du langage naturel pour ensuite extraire des enchaînements fréquents de voisinages de mots. Les premières analyses de ces motifs sont riches d'enseignement et les découvertes non dirigées, parfois surprenantes, mènent à des réflexions très pertinentes. Nos travaux peuvent ainsi compléter les analyses antérieures qui développaient l'idée d'une unité d'esprit au sein du *Lancelot*[28].

Les perspectives de ces travaux sont doubles. Tout d'abord, le travail d'analyse doit être poursuivit en particulier pour traiter les motifs propres à la thématique de la parenté, pour traiter le corpus dans sa globalité et mettre en évidence des tendances au sein des différents documents. Ensuite, il est possible d'adopter des algorithmes de fouille de données intégrant des contraintes supplémentaires afin de raffiner les motifs ou d'explorer plus en avant une piste envisagée.

Références

Brill E. (1994). « Some advances in rule-based part of speech tagging ». In *Proceedings of the Twelfth National*

[28] Ferdinand Lot, *Etude sur le Lancelot en Prose*, op. cit. p. 65 à 107.

Conference on Artificial Intelligence (AAAI-94), pp 722-727.

Cornuéjols A., Miclet L. (2002). « *Apprentissage artificiel, Concepts et algorithmes* », Eyrolles.

Dupuis F., Lebart L. (2008). « Visualisation, validation et sériation. Application à un corpus de textes médiévaux ». In *Proceedings of JADT'08 (Journées internationales d'Analyse statistique des Données Textuelles)*, pp 433-444.

Heiden S., Guillot C. (2003). « Capitalisation des savoirs par web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval ». In *Ancien et moyen français sur le Web, enjeux méthodologiques et analyse du discours*, Pierre Kunstmann, France Martineau & Danielle Forget Eds. Les éditions David.

Heiden S. (2004). « Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex ». In *Proceedings of JADT'04 (Journées internationales d'Analyse statistique des Données Textuelles)* "Le poids des mots", pp 577-588.

Heiden S., Lavrentiev A. (2004). « Ressources électroniques pour l'étude des textes médiévaux : approches et outils ». In *Revue Française de Linguistique Appliquée*, IX(1), Linguistique et informatique : nouveaux défis, B. Habert (resp.).

Jacquemin C., Bourigault D. (2003). « Term Extraction and Automatic Indexing », In *The Oxford Handbook of Computational Linguistics*, Mitkov R. (ed), Oxford University Press, pp. 599-615.

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

Masseglia F. , Teisseire M., Poncelet P. (2004). « Recherche des motifs séquentiels », In Revue *Ingénierie des Systèmes d'Information* (ISI), numéro spécial "*Extraction de motifs dans les bases de données*". Décembre 2004, Vol. 9, N° 3-4, pp. 183-210.

Mayaffre D. (2007). « L'analyse de données textuelles aujourd'hui : du corpus comme une urne, au corpus comme un plan. Bilan sur les travaux actuels de topographie/topologie textuelle ». In *Lexicométrica* thématique.

Prévost S. et Heiden S. (2004) « Etiquetage d'un corpus hétérogène de français médiéval : enjeux et modalités ». *1st Freiburg Workshop on Romance Corpus Linguistics, Freiburg, in : Romanistische Korpuslingustik: Korpora und gesprochene Sprache, Romance Corpus Linguistics: Corpora and Spoken Language* p. 127-136.

Rabatel J., Lin Y., Pitarch Y., Saneif H., Serp C., Roche M., Laurent A. (2008). « Visualisation des motifs séquentiels extraits à partir d'un corpus en Ancien Français », In *Proceedings of EGC'08* (session démonstration), p237-238.

Schmid, H. (1994). « Probabilistic Part-of-Speech Tagging UsingDecision Trees », In *Proceedings of the International Conference on NewMethods in Language Processing*, pp. 44–49.

Serp C., Cazal E., Laurent A., Roche M. (2008). « TERVOTIQ : un système de vote pour l'extraction de la terminologie d'un corpus en français médiéval », In *Proceedings of JADT'08 (Journées internationales d'Analyse statistique des Données Textuelles)*, pp. 1069-1020.

Stein A. (2003). « Part of Speech Tagging and Lemmatisation of Old French Texts », *http://www.unistuttgart.de/lingrom/stein/forschung/altfranz/afrlemma.pdf*.

Remerciements

Nous tenons à remercier M. Max Engammare, directeur de la librairie Droz, pour son aimable autorisation d'utiliser une version numérique du *Lancelot*.