

Extraction de terminologie pour l'ancien français : la quête du Graal

Emmanuel Cazal*, Claire Serp**, Mathieu Roche*, Anne Laurent*

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

**Université Montpellier 3, serpclaire@yahoo.fr

Résumé. La fouille de textes trouve de très nombreuses applications (annotation automatique, identification d'auteurs, segmentation thématique, etc.). Des méthodes efficaces existent aujourd'hui pour extraire par exemple la terminologie comme support du processus d'extraction. Cependant si les méthodes sont valides pour des textes rédigés en français, anglais etc, il n'en est pas de même pour les textes en ancien français. Ceux-ci, trop complexes pour être traités par les méthodes automatiques existantes, restent en effet encore exclusivement étudiés par les experts du domaine (spécialistes en ancien français, médiévistes) avec des méthodes manuelles. Nous étudions donc dans cet article l'une des premières étapes du traitement de textes associée à l'extraction de la terminologie en considérant un corpus en ancien français. Nous présentons les principaux problèmes rencontrés et les limites d'une approche classique dans un tel contexte.

1 Introduction

La fouille de textes est un domaine de recherche très actif et de nombreuses propositions ont été réalisées ayant toujours comme objectif un traitement le plus automatique possible, afin de dédier les interventions de l'expert à l'évaluation des connaissances extraites et non au moyen de les obtenir. Certains de ces travaux se basent sur une extraction de la terminologie comme support du processus d'extraction et des résultats très prometteurs ont déjà été obtenus (Azé et Roche (2003); Roche et al. (2004b)). Notons qu'avant d'extraire la terminologie, des étapes préliminaires d'acquisition d'un corpus et de normalisation de ce dernier sont nécessaires (étapes non explicitement décrites dans cet article). Précisons enfin qu'à l'heure actuelle, il semble important de placer l'expert au cœur du processus de fouille de textes, voie dans laquelle ces travaux s'inscrivent.

Le traitement proposé ici se situe dans un contexte particulier qui le rend complexe selon deux aspects. Premièrement les textes manipulés sont en ancien français et rares sont les travaux sur de tels documents. Ceux-ci, trop complexes pour être traités par les méthodes automatiques existantes, restent en effet encore exclusivement étudiés par les experts du domaine (spécialistes en ancien français, médiévistes) avec des méthodes manuelles. Deuxièmement, la base de données correspond à un volume important de textes. En effet, le corpus étudié comprend plus de deux mille pages réparties en deux grands ensembles, le cycle Lancelot-Graal (5 ouvrages) et le Perlesvaus. L'ancien français pose ici deux problèmes majeurs. Tout d'abord, comme le latin dont elle est issue, c'est une langue à déclinaison, c'est-à-dire que les mots

en ancien français portent des marques particulières en fonction de leur place dans la phrase (par exemple, le mot chevalier au singulier en position de sujet s'écrit avec un S, tandis que le même mot en complément d'objet direct s'écrit sans S). La deuxième particularité de cette langue est qu'elle n'a pas de normes orthographiques "fixes", les écrivains utilisant différentes formes pour un même mot, et cela au sein d'un même texte (nous pouvons pas exemple citer le mot soeur, que l'on peut trouver dans le tome VII du Lancelot sous les formes suivantes : soeur, serours, seur, seror, seurs , suer). Dès lors, il paraît évident qu'un lexique, aussi complet soit-il, ne peut intégrer toutes les variantes orthographiques d'un même mot, et doit se limiter à répertorier les formes les plus fréquentes.

Les recherches menées sont issues d'une première étude d'un sous-ensemble du corpus. Basée sur le relevé d'occurrences, cette étude est réalisée dans le cadre d'une thèse en littérature médiévale sur le thème "Filiation, identité et problèmes de parenté dans les romans du Graal en prose". Celle-ci a permis de mettre en évidence des différences importantes dans le traitement de l'imaginaire de la parenté d'un texte à l'autre, notamment grâce à l'étude du contexte dans lequel apparaissait le terme de "frère". Il est alors apparu important de réaliser une étude plus systématique basée sur une méthode appropriée d'extraction de connaissances dans l'objectif d'automatiser l'identification de corrélations possibles. Le problème soulevé alors est le suivant : dans quelle mesure la démarche classique de fouille de textes peut-elle s'adapter à un corpus en ancien français ? Les résultats seront-ils aussi intéressants que ceux obtenus sur des documents plus modernes ? Tel est l'objet de cet article qui présente en section 2 le processus adopté : choix d'une méthode d'extraction de la terminologie, choix d'un étiqueteur, choix des lexiques pour ensuite souligner les problèmes des lexiques utilisés. Section 3, nous détaillons le protocole d'évaluation ainsi que les différentes expérimentations menées. Enfin, en conclusion, nous dressons les nombreuses perspectives associées à ce travail.

2 Démarche globale et choix de réalisation

2.1 Extraction de la terminologie

De multiples approches de recherche terminologique ont été développées afin d'extraire les termes pertinents à partir d'un corpus. Les méthodes d'extraction de la terminologie sont fondées sur des méthodes statistiques et/ou syntaxiques. Le système TERMINO de David et Plante (1990) est un outil précurseur qui s'appuie sur une analyse syntaxique afin d'extraire les termes nominaux. Cet outil effectue une analyse morphologique à base de règles, suivie de l'analyse des termes nominaux à l'aide d'une grammaire. Les travaux de Smadja (1993) (XTRACT) s'appuient sur une méthode statistique. XTRACT extrait, dans un premier temps, les termes binaires situés dans une fenêtre de dix mots. Les termes binaires sélectionnés sont ceux qui dépassent d'une manière statistiquement significative la fréquence due au hasard. L'étape suivante consiste à extraire les termes plus généraux (termes de plus de deux mots) contenant les termes binaires trouvés à la précédente étape. ACABIT de Daille (1994) effectue une analyse linguistique afin de transformer les termes nominaux en termes binaires. Ces derniers sont ensuite triés selon des mesures statistiques. Contrairement à ACABIT qui est fondé sur une méthode statistique, LEXTER de Bourigault (1993) et SYNTAX de Bourigault et Fabre (2000) s'appuient essentiellement sur une analyse syntaxique afin d'extraire la terminologie du domaine. La méthode consiste à extraire les groupes nominaux maximaux. Ces groupes (appe-

lés syntagmes) sont alors décomposés en termes de “têtes” et d’“expansions” à l’aide de règles grammaticales. Les termes sont alors proposés sous forme de réseau organisé en fonction de critères syntaxiques.

Dans nos travaux, nous utiliserons le système EXIT (Roche et al. (2004c)). À partir d’un corpus étiqueté grammaticalement (voir section suivante), le système permet d’extraire les candidats termes respectant des patrons syntaxiques définis (Nom-Nom, Nom-Préposition-Nom, etc.). Ces candidats termes sont alors classés en utilisant diverses mesures statistiques (Roche et al. (2004a)). Une des caractéristiques de ce système est son aspect itératif permettant d’effectuer une nouvelle recherche terminologique à partir du corpus avec prise en compte de la terminologie du domaine acquise aux étapes précédentes. Les divers paramètres adaptés à cette approche itérative permettent alors de détecter des candidats termes pertinents (appelés termes) très spécifiques c’est-à-dire composés de plusieurs mots.

Comme nous l’avons présenté dans cette section, pour pouvoir extraire la terminologie, il est nécessaire d’effectuer un étiquetage grammatical du corpus. L’étiquetage permet alors d’extraire les candidats termes respectant un patron défini (Nom-Nom, Nom-Préposition-Nom, Nom-Adjectif, Adjectif-Nom, etc.). La section suivante présente les systèmes d’étiquetage grammatical les plus utilisés.

2.2 Étiquetage grammatical

Le TreeTagger de Schmid (1994) estime la probabilité qu’un mot ait une étiquette grammaticale (Nom, Adjectif, Déterminant, etc.) en s’appuyant sur des arbres de décision binaires (Quinlan (1986)). Ces derniers sont construits récursivement à partir d’un ensemble de trigrammes connus (suites de trois étiquettes grammaticales consécutives constituant l’ensemble d’apprentissage). Le processus complet de construction des arbres de décision est décrit dans les travaux de Schmid (1994).

L’étiqueteur de Brill appose une étiquette grammaticale à chacun des mots d’un texte en utilisant un lexique, des règles lexicales et des règles contextuelles. Dans l’approche développée dans les travaux de Brill (1994), l’auteur s’appuie sur un corpus d’apprentissage du *Wall Street Journal*. Le but est alors d’apprendre des règles d’étiquetage à partir de ce corpus. Ce corpus est annoté manuellement et représente l’ensemble des annotations justes (*truth*). À chaque étape d’apprentissage, des règles sont modifiées et le résultat de l’étiquetage avec ces nouvelles règles est comparé avec le corpus représentant l’ensemble des annotations justes. Tant qu’un nombre d’erreurs seuil dans l’étiquetage subsiste, le processus d’apprentissage continue. Les transformations des étiquettes s’effectuent (1) en changeant une étiquette par une autre suivant les mots ou les étiquettes des mots proches, (2) en utilisant certaines caractéristiques pour les mots inconnus (lettres en majuscules pour les noms propres, suffixe des mots, etc.). Dans une série d’expérimentations sur un corpus du *Wall Street Journal*, les résultats donnent une précision de 96.5% sur l’ensemble du corpus test. Des systèmes tels que ETIQ (Amrani et al. (2004)) permettent d’ajouter des règles lexicales aux mots inconnus de l’étiqueteur de Brill. ETIQ permet également d’ajouter aisément des règles contextuelles. Par ailleurs, des méthodes d’apprentissage peuvent être utilisées pour proposer de nouvelles règles à l’expert.

Les travaux présentés dans cet article s’appuient sur l’utilisation de l’étiqueteur de Brill pour pouvoir extraire la terminologie du domaine. N’ayant pas de corpus étiquetés manuellement en relation directe avec le corpus spécialisé étudié, nous ne pouvons mettre en œuvre une phase d’apprentissage supervisé comme dans les travaux de Stein (2003). Dans un premier

temps, notre approche consiste à construire un lexique adapté au corpus étudié. La méthode mise en place pour construire ces lexiques qui seront utilisés par l'étiqueteur de Brill est détaillée dans la section suivante. Notre approche est généralisable et peut prendre en compte toute nouvelle ressource propre à l'ancien français (par exemple, des ressources issues du Tree Tagger mises à jour en octobre 2006¹). L'extraction de la terminologie en utilisant EXIT pourra alors être effectuée à partir de ce corpus étiqueté.

Utilisation d'un lexique en ancien français pour l'étiquetage grammatical. Afin de réaliser un étiquetage de bonne qualité des textes en ancien français, nous utilisons deux lexiques. Le premier lexique, en français moderne, contient plus de 440 000 mots, obtenu auprès de l'INaLF (Institut National de la Langue française²). Le second lexique, en ancien français, contient, après prétraitements, un peu plus de 45 000 mots. Dans chacun de ces lexiques chaque mot est associé à une ou plusieurs étiquettes. Malheureusement la première difficulté vient de la structure des lexiques qui est différente. En effet, le lexique en français moderne possède, pour chaque ligne, la structure suivante : *mot étiquette1 étiquette2*. Un exemple nous donne : *habitacle Nom_singulier*. Le lexique en ancien français possède quant à lui, pour chaque ligne, la structure suivante : "*mot*", "*étiquetteA*", "*étiquetteB / étiquetteC*" (par exemple "*abitacle*", "*Nom_singulier_masculin/Nom_singulier_féminin*"). Ce dernier est issu des travaux de Mr Douglas C. Walker de l'Université de Calgary et il est disponible à l'URL : <http://www.acs.ucalgary.ca/~dcwalker/Dictionary/dict.html>. Ces deux lexiques seront par la suite désignés par lexique AF pour le lexique en ancien français et par lexique FM pour le lexique en français moderne.

Notre objectif est d'améliorer la qualité de notre étiquetage en augmentant dans le lexique le nombre de mots couvrant le domaine. C'est pourquoi nous avons fusionné les deux lexiques (AF et FM) afin de s'assurer d'un étiquetage grammatical de qualité satisfaisante. Nous avons adopté la structure suivante : (i) une seule occurrence par ligne, (ii) pour chaque ligne, conserver les étiquettes correspondant au mot en ancien français ou au mot en français moderne. Nous obtenons donc pour chacune des lignes la structure suivante : *mot étiquette1 ... étiquetteX*. Cela nous donne le format suivant, pour les mots *utilisateur* et *germain* : pour le français moderne : *utilisateur Nom Adjectif*, et pour l'ancien français : *germain Adjectif Nom_masculin*.

Les prétraitements cités plus haut consistaient à rendre le formalisme du lexique en ancien français identique à celui du lexique en français moderne. L'exemple *germain Adjectif Nom_masculin* avait, avant prétraitements, le format suivant : "*germain*", "*adj/sm*". Les étapes de prétraitements devaient donc permettre d'identifier les lignes contenant un seul ou plusieurs mots eux-mêmes associés à une ou plusieurs étiquettes. L'adaptation des étiquettes du lexique AF au format du lexique FM a été réalisée au travers d'une liste de correspondance établie en commun avec la médiéviste co-auteur de cet article. Ceci montre que pour l'étude de données complexes, dans notre cas l'ancien français, une étroite collaboration avec l'expert est essentielle.

Une fois les étapes de prétraitements terminées, il reste à fusionner les occurrences des lexiques AF et FM afin de constituer un seul lexique mixte. Les différentes conditions utilisées permettent de traiter, pour chaque mot, les trois cas de figure rencontrés, c'est-à-dire : (i) les

¹<http://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html>

²<http://www.inalf.cnrs.fr>

mots du lexique AF (mot AF) appartiennent aussi au lexique FM, (ii) les mots du lexique AF n'appartiennent pas au lexique FM, (iii) les mots du lexique FM (mot FM) n'appartiennent pas au lexique AF. En effet, afin d'assurer la cohérence entre le lexique mixte et les corpus de données à traiter nous privilégions les entrées du lexique AF. Une entrée ou "occurrence" est une ligne d'un lexique contenant le mot et son (ses) étiquette(s).

L'algorithme mettant en œuvre les conditions utilisées pour réaliser la fusion se compose donc de trois étapes. La *première étape* consiste à vérifier l'existence de chacune des occurrences du lexique AF dans le lexique FM. Si les mots sont présents dans les deux lexiques, les entrées du lexique AF seront écrites dans le lexique mixte. Notons que les étiquettes communes à AF et FM sont positionnées dans l'ordre de FM dans le lexique mixte. En effet, les connaissances relatives à cet ordre qui sont disponibles au niveau du lexique FM (et non au niveau du lexique AF) sont importantes pour privilégier les premières étiquettes lors de la phase d'étiquetage. Par exemple, le mot AF "musique Adjectif Nom_singulier" et le mot FM "musique Nom_singulier" co-existent, par conséquent l'occurrence de sortie sera "musique Nom_singulier Adjectif". La *deuxième étape* ajoute, dans le lexique mixte, toutes les entrées du lexique AF qui ne sont pas dans le lexique FM. Puis, la *troisième étape* vérifie que toutes les occurrences du lexique FM sont ajoutées au lexique mixte, c'est-à-dire que nous complétons le lexique mixte des mots du lexique FM qui ne seraient pas dans le lexique AF.

Résultats de la fusion des lexiques. Afin d'implémenter cette méthode, un programme en Perl a été spécifiquement développé. Le lexique AF contenait 2138 mots qui existaient aussi dans le lexique FM ce qui représente 4,72% de mots en double. 43083 entrées du lexique AF étaient quant à elles inconnues du lexique FM, ce qui représente 95,28%. Ceci signifie que les 43083 entrées qui n'apparaissent pas dans le lexique FM ont été intégrées dans le lexique mixte. Nous pouvons dire que cette fusion des lexiques était utile et nous regretterons le fait que le lexique AF ne représente qu'un dixième du lexique FM en terme de nombres d'entrées.

L'étude du nombre d'entrées du lexique FM montre qu'il contient beaucoup de déclinaisons verbales. Par exemple, pour le verbe *garder*, il y a dans le lexique FM cinq déclinaisons contre une seule pour le lexique AF. C'est pour cela que nous augmentons le nombre d'entrées du lexique mixte en déclinant les verbes en ancien français grâce à un traitement spécifique.

2.3 Traitement des verbes dans le processus d'étiquetage grammatical

Comme nous l'avons vu, le nombre de mots en ancien français dans notre lexique mixte est faible, c'est pour cela que nous avons décidé d'utiliser une méthode de déclinaisons des verbes. Cette méthode implique de disposer de plusieurs types d'étiquettes. Nous utilisons l'étiqueteur grammatical de Brill (1994) qui peut gérer suffisamment de variantes d'étiquettes pour qu'il nous soit possible de préciser la forme de conjugaison d'un verbe. Par exemple, pour le verbe *ochire* (tuer) avec comme base verbale *ochi*, nous identifions dans le corpus les mots : *ochire* qui sera étiqueté VNCF, c'est-à-dire verbe non conjugué à l'infinitif, *ochiroit* qui sera étiqueté VCJ :sg, c'est-à-dire verbe conjugué, *ochistrent* qui sera étiqueté VCJ :sg, verbe conjugué.

Nous exploitons cette gestion des étiquettes et donc des déclinaisons verbales pour augmenter le nombre de mots en ancien français dans le lexique mixte. La méthode retenue repose sur l'utilisation d'une liste de 36 bases verbales, établie par la médiéviste ainsi que sur l'uti-

lisation de la liste des mots du corpus non couverts par le lexique mixte. Nous avons retenu cette liste de mots du corpus non couverts afin d'augmenter, dans le lexique mixte, le nombre d'entrées propres au type de corpus étudié.

De la comparaison de ces deux listes, nous obtenons tous les nouveaux mots étiquetés grâce aux bases verbales utilisées. Une fois cette nouvelle liste validée par l'experte, nous enrichissons le lexique mixte. Cette méthode nous a permis d'identifier 110 mots du corpus qui n'étaient pas dans le lexique mixte. Le processus complet est illustré par la figure 1.

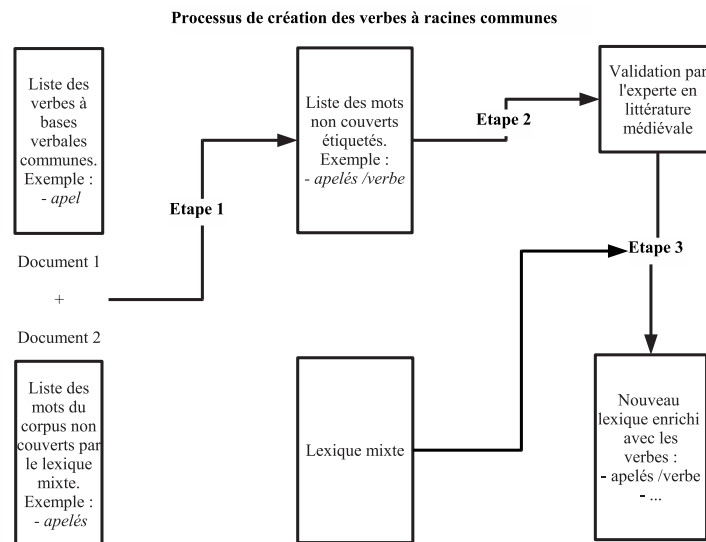


FIG. 1 – *Processus de création des verbes à bases verbales communes.*

La figure 1 reprend, étape par étape, le processus d'enrichissement du lexique mixte par le traitement des déclinaisons verbales. La première étape consiste à comparer le fichier contenant la liste des bases verbales au fichier contenant la liste des mots du texte qui ne sont pas identifiés dans le lexique mixte. Cette comparaison met en évidence tous les mots qui possèdent une des bases verbales. La seconde étape est une étape de validation par l'expert de la fiabilité des résultats obtenus. L'étape finale consiste à intégrer dans le lexique mixte les résultats de l'étape 2. Ainsi, en fonction des résultats que nous retourne ce traitement des verbes du corpus, nous affinons notre méthode.

La section suivante présente une étude quantitative et qualitative de notre approche.

3 Expérimentations

Deux types d'expérimentations sont mises en œuvre pour évaluer la qualité du lexique mixte créé. Les premières expérimentations consistent à effectuer une étude quantitative en mesurant le taux de couverture. Les expérimentations suivantes s'appuient sur une étude qualitative par l'expertise manuelle de l'étiquetage obtenu en utilisant des lexiques.

3.1 Évaluation quantitative : calcul du taux de couverture

Le nombre d'entrées dans le lexique a bien entendu des conséquences sur la qualité de l'étiquetage. En effet, plus les mots du texte à étiqueter sont présents dans le lexique et plus la qualité de l'étiquetage devrait être améliorée. Nous proposons de calculer le taux de couverture des mots du texte en utilisant différents lexiques. Deux types de calculs sont proposés. La première méthode consiste à prendre seulement en compte les mots uniques. Ainsi, un même mot répété plusieurs fois dans le corpus ne sera comptabilisé qu'une seule fois dans le calcul de ce taux de couverture. Ce dernier, donné par la formule (1), a pour objectif de donner moins de poids aux mots fréquents (prépositions, déterminants, etc.) qui ne sont pas nécessairement représentatifs pour le domaine d'étude.

$$Tdc_{unique} = \frac{\text{nombre de mots uniques du texte présents dans le lexique}}{\text{nombre de mots uniques du texte}} \quad (1)$$

Le second calcul du taux de couverture donné par la formule (2) prend en compte tous les mots du corpus. Ainsi, le nombre de fois où un mot apparaît dans le corpus sera pris en compte dans le calcul de ce taux de couverture appelé Tdc_{global} . Dans le cas où les mots fréquents sont présents dans le lexique, Tdc_{global} aura alors une valeur bien supérieure à Tdc_{unique} . L'objectif ici est d'évaluer si les lexiques utilisés couvrent les mots principaux de l'ancien français.

$$Tdc_{global} = \frac{\text{nombre de mots du texte présents dans le lexique}}{\text{nombre de mots du texte}} \quad (2)$$

Le tableau 1 présente un fragment de texte en ancien français issu du corpus et les mots en **gras** sont les mots contenus dans le lexique mixte.

Li soumiens estoit moult bien cargiés de joiaus et de vaselemente et de deniers.
--

TAB. 1 – Fragment de texte écrit en ancien français.

Dans cet exemple, nous pouvons identifier : le nombre de mots total dans le texte (14), le nombre de mots uniques du texte (11), le nombre de mots uniques du texte contenus dans le lexique mixte (8), le nombre total de mots du texte contenus dans le lexique mixte (11). Nous pouvons donc établir les taux de couverture suivants : $Tdc_{unique} = \frac{8}{11}$ donc $Tdc_{unique} = 72\%$ et $Tdc_{global} = \frac{11}{14}$ donc $Tdc_{global} = 78\%$. Ces résultats montrent que l'efficacité de notre lexique dépend de son importance en terme de nombre d'entrées. Le calcul fondé sur le nombre de mots uniques (Tdc_{unique}) permet de représenter la quantité d'entrées applicables à un texte donné. Le calcul de l'ensemble des mots du corpus (Tdc_{global}) montre le poids que certains mots peuvent avoir grâce à leur répétition dans le texte. Le calcul de la couverture en utilisant différents lexiques est donné dans le tableau 2.

Le tableau 2 montre que, sans les bases verbales, les lexiques en AF et en FM couvrent presque de façon identique le corpus. En effet, pour le lexique AF et le lexique FM nous avons 35% et 37% pour le Tdc_{unique} et 68% et 70% pour le Tdc_{global} . L'évolution des Tdc_{unique} des lexiques AF et mixte, après l'ajout des bases verbales (respectivement 35% versus 38% et 53% versus 56%) montre l'intérêt de notre opération. Notons qu'entre le lexique mixte et le lexique mixte avec bases verbales, les valeurs du Tdc_{global} sont sensiblement les mêmes. Cela s'explique par la faible répétition de certains mots dans le corpus.

	Lexiques				
	AF	FM	mixte	AF avec verbes	mixte avec verbes
TdC_{unique}	35%	37%	53%	38%	56%
TdC_{global}	68%	70%	81%	71%	82%

TAB. 2 – Taux de couverture.

3.2 Évaluation qualitative de l'étiquetage grammatical

Après avoir évalué quantitativement notre méthode de construction du lexique mixte, une évaluation manuelle qualitative est présentée dans cette section. À partir de cette évaluation, le taux d'erreur est calculé. Ce dernier représente la proportion d'étiquettes erronées parmi les étiquettes appliquées. Ainsi, une évaluation manuelle d'un sous-ensemble du corpus du Lancelot représentant 171 mots a été mise en œuvre (voir tableau 3).

Lexiques utilisé pour l'étiquetage	Taux d'erreur de l'étiquetage
ancien français	46%
français moderne	63%
mixte avec verbes	35%

TAB. 3 – Taux d'erreur de l'étiquetage grammatical.

Les résultats du tableau 3 montrent que le taux d'erreur est beaucoup plus faible avec le lexique mixte comparativement aux lexiques en ancien français et en français moderne. Notre approche se révèle donc très encourageante et pourrait être adaptée en utilisant de nouvelles connaissances comme les ressources issues du Tree Tagger. Le taux d'erreurs assez important (63%) obtenu avec le lexique en français moderne montre que le vocabulaire utilisé dans le corpus du Lancelot est très spécifique. La raison pour laquelle le taux d'erreur trouvé avec le lexique en ancien français est assez important (46%) s'explique par le fait que la quantité de données du lexique n'est pas encore suffisante, notamment en terme de déclinaisons verbales.

L'utilisation des différentes connaissances des deux lexiques diminue significativement le taux d'erreur qui reste pourtant assez important. Plusieurs raisons peuvent expliquer un tel résultat. Outre l'absence de certains mots du lexique qui provoque des erreurs, l'association de plusieurs étiquettes possibles pour un même mot peut également provoquer des résultats erronés. Par exemple, en français moderne, le mot "entrée" peut être à la fois un nom ou un participe passé. Des problèmes similaires se posent en ancien français qui provoquent des ambiguïtés lors de la phase d'étiquetage grammatical.

Par ailleurs, l'étiqueteur de Brill n'utilise pas seulement un lexique mais également des règles lexicales et contextuelles. En particulier, dans certains cas, ces dernières peuvent être très différentes pour les deux types de textes (ancien français et français moderne). Ceci a alors provoqué des erreurs au niveau de l'étiquetage du corpus du Lancelot. Une prochaine étape de nos travaux consistera à adapter les règles lexicales et contextuelles aux corpus spécifiques que nous étudions.

Extraction de la terminologie. Nous estimons qu'avec un taux d'erreurs de 35% concernant l'étiquetage grammatical, l'extraction de la terminologie n'est pas encore une étape à mener à

grande échelle. En effet, les premiers tests d'extraction de la terminologie ont montré que les termes obtenus sont, en général, de qualité peu satisfaisante. Les erreurs constatées sont dues à l'étape d'étiquetage grammatical qui n'est pas encore aboutie. Notons que l'utilisation du processus global de fouille de textes est primordiale pour mettre en relief des erreurs rencontrées aux étapes précédentes. En effet, un tel processus nécessite d'avoir des étapes d'actions mais également de rétro-actions pour améliorer le traitement global appliqué. Ainsi, l'étape de terminologie permet de relever des erreurs non seulement au niveau de l'étiquetage mais potentiellement lors de la phase de normalisation du corpus voire au niveau de son acquisition.

4 Conclusion

La problématique présentée ici vise à rechercher des informations nouvelles dans un corpus écrit en ancien français. Pour cela, nous proposons de mettre en œuvre un processus global de fouille de textes, comprenant une étape d'acquisition du corpus suivie d'une tâche de normalisation. À partir du corpus normalisé, une phase d'extraction de la terminologie est mise en œuvre, nécessitant au préalable l'étiquetage grammatical des textes. Cet article s'intéresse plus spécifiquement à la phase d'extraction de la terminologie et se focalise sur l'étape préalable d'étiquetage grammatical des textes. Celle-ci s'effectue en appliquant l'étiqueteur de Brill qui utilise un lexique. Cet article présente une méthode pour constituer un lexique (appelé lexique mixte) adapté à l'ancien français. Les résultats présentés s'appuient sur l'utilisation du lexique mixte et donnent des résultats encourageants quant à la qualité de l'étiquetage obtenu. Cependant, des erreurs d'étiquetage subsistent, essentiellement dues à des entrées non présentes dans le lexique et au traitement encore partiel des déclinaisons verbales. La prise en compte des règles lexicales et contextuelles adaptées à l'ancien français devrait également améliorer significativement les résultats obtenus. Ces erreurs d'étiquetage ont une conséquence immédiate : l'extraction de la terminologie proprement dite n'a pas pu être menée. Le Graal n'est donc pas encore à notre portée ! Précisons enfin que la méthode généralisée proposée ici ne nécessite pas de phase d'apprentissage, ce qui aurait été très coûteux (nécessite un grand corpus totalement étiqueté). Une telle approche a été menée dans des travaux relatifs à des textes écrits en ancien français (Stein (2003)). Une perspective à notre travail consiste à comparer l'utilisation de notre lexique mixte et celui de Stein (2003).

Références

- Amrani, A., Y. Kodratoff, et O. Matte-Tailliez (2004). A semi-automatic system for tagging specialized corpora. In *Proceedings of PAKDD'04*, pp. 670–681.
- Azé, J. et M. Roche (2003). Une application de la fouille de textes : l'extraction des règles d'association à partir d'un corpus spécialisé. *Revue RIA-ECA numéro spécial EGC03 17*, 283–294.
- Bourigault, D. (1993). Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *T.A.L.* 34(2), 105–118.
- Bourigault, D. et C. Fabre (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires* 25, 131–151.

- Brill, E. (1994). Some advances in transformation-based part of speech tagging. In *AAAI, Vol. 1*, pp. 722–727.
- Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph. D. thesis, Université Paris 7.
- David, S. et P. Plante (1990). De la nécessité d'une approche morpho syntaxique dans l'analyse de textes. In *Intelligence Artificielle et Sciences Cognitives au Québec*, Volume 3, pp. 140–154.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning 1*, 81–106.
- Roche, M., J. Azé, Y. Kodratoff, et M. Sebag (2004a). Learning interestingness measures in terminology extraction. A ROC-based approach. In *Proceedings of "ROC Analysis in AI" Workshop (ECAI 2004), Valencia, Spain*, pp. 81–88.
- Roche, M., J. Azé, O. Matte-Tailliez, et Y. Kodratoff (2004b). Mining texts by association rules discovery in a technical corpus. In *Proceedings of IIPWM'04 (Intelligent Information Processing and Web Mining), Springer Verlag series "Advances in Soft Computing"*, pp. 89–98.
- Roche, M., T. Heitz, O. Matte-Tailliez, et Y. Kodratoff (2004c). EXIT : Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. In *Proceedings of JADT'04*, Volume 2, pp. 946–956.
- 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.
- Smadja, F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistics 19*(1), 143–177.
- Stein, A. (2003). Part of speech tagging and lemmatisation of old french texts. In <http://www.uni-stuttgart.de/lingrom/stein/forschung/altfranz/aflemma.pdf>.

Summary

Text mining is an active research area with numerous applications (automatic annotation, author recognition, thematic clustering, ...). Efficient methods have been designed to tackle the problem of extracting the terminology. This terminology can then be used for knowledge discovering. However, although these methods are valid on texts written in French, English, etc, they are not suitable for texts written in Old French. These texts are indeed too complex to be treated as classical texts. Thus they are currently studied manually by human experts. For this reason, we address here the problem of automatically extracting the terminology from old French texts. This process is meant as then being integrated in a knowledge discovery process in future work. In this paper, we detail our approach, together with the problems we encountered on this very first step and the limits of classical limits on such complex data.