TERVOTIQ : Un système de vote pour l'extraction de la terminologie d'un corpus en français médiéval

Claire Serp, Emmanuel Cazal, Anne Laurent, Mathieu Roche

TERVOTIQ : un système de vote pour l'extraction de la terminologie d'un corpus en français médiéval

Claire Serp[1], Emmanuel Cazal[2], Anne Laurent[2], Mathieu Roche[2]

[1]Université Montpellier 3, serpclaire@yahoo.fr

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

Abstract

Automatic terminology extraction from texts is a very useful task for users like researchers in literature. Tools proposed to perform this extraction require texts tagged with grammatical categories. When dealing with texts written in Old French, this tagging is not an easy task. In this paper, we thus address the problem of tagging such texts. We consider the combination of two existing tools for tagging words: TreeTagger and Brill. We propose the so-called TERVOTIQ system (TERminologie par VOte selon l'éTIQuetage), which aims at enhancing the quality of the terminology extracted by improving the tagging. This tool is based on the use of a criterion of reliability in order to consider those terms that were extracted from the texts treated by the two taggers. Our approach allows the ordering of terms by the value of the criterion of reliability (the higher value of the criterion represents the more relevant term), and by the number of occurrences. Experiments have shown that the terms that we extract using our tool are more relevant than the ones extracted without using the combination of the other two taggers.

Keywords: terminology, tagging, Old French.

Résumé

L'extraction automatique de terminologie dans les textes est une tâche utile pour bon nombre d'utilisateurs finaux tels que les chercheurs en littérature. Des outils effectuant cette extraction utilisent des textes étiquetés selon des catégories grammaticales. Dans ce contexte, le traitement des textes en français médiéval est une tâche complexe, dans la mesure où l'orthographe est fluctuante et qu'il n'existe donc pas de lexique comprenant toutes les flexions d'un même mot. Dans cet article, nous nous intéressons au problème de l'étiquetage de tels textes. Nous considérons la combinaison de deux outils permettant un étiquetage grammatical : TreeTagger et Brill. Nous proposons le système TERVOTIQ (TERminologie par VOte selon l'éTIQuetage) qui a pour objectif d'accroître la qualité de la terminologie extraite grâce à l'amélioration de l'étiquetage. Ce système fondé sur l'utilisation d'un indice de fiabilité permet de privilégier les termes communs (termes extraits à partir des textes traités par plusieurs étiqueteurs). Notre approche permet de classer les termes en fonction de l'indice le plus fiable (la plus grande valeur du critère représente le terme le plus pertinent) puis selon le nombre d'occurrences. Les expérimentations ont montré que les termes extraits avec notre système sont plus pertinents que ceux extraits sans utiliser de combinaison des deux étiqueteurs.

Mots-clés : terminologie, étiquetage, français médiéval.

1. Introduction

Les travaux présentés dans cet article reposent sur l'extraction de la terminologie à partir de corpus en ancien français. Le corpus étudié comprend plus de deux milles pages réparties en deux grands ensembles, le cycle Lancelot-Graal (5 ouvrages) et le Perlesvaus. Ce corpus est constitué de 706 938 mots. Soulignons que les expérimentations menées dans cet article s'appuient sur un sous-ensemble du corpus composé de 11 725 mots.

L'ancien français pose deux problèmes majeurs lorsque l'on souhaite en avoir une vision d'ensemble. Tout d'abord, comme le latin dont elle est issue, c'est une langue à déclinaisons. C'est-à-dire que bien que le système soit plus simple que le latin, 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 », ce qui veut dire que les écrivains utilisent 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, serours, 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 usitées.

Les recherches menées sont issues d'une première étude d'un sous-ensemble du corpus. Cette étude s'appuie sur le relevé d'occurrences qui a été réalisé 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 » dont certains résultats ont été récemment communiqués (Serp, 2007). 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 mettre en oeuvre un processus d'extraction de la terminologie adapté à l'ancien français afin d'aider à cette analyse. Des résultats préliminaires relatifs à l'extraction de la terminologie à partir de ce corpus ont été présentés dans (Cazal et al., 2007). Par ailleurs, des outils de visualisation des textes en ancien français associés à des tâches de fouille de données actuellement menées ont également été développés (Rabatel et al., 2008).

Parmi les travaux sur le traitement de textes en ancien français, nous pouvons citer ceux de la BFM1 (Base de Français Médiéval) fondés sur le système *Weblex*[2] développé au laboratoire ICAR de l'Ecole Normale Supérieure de Lettres et Sciences Humaines (Heiden, 2004). Les textes sont tout d'abord normalisés selon le format XML[3] permettant d'obtenir le document selon une structure arborescente avec les méta-données associées. Ensuite, la recherche d'information, à l'aide d'un moteur de recherche, offre 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.

Dans notre article qui propose une étude qui s'appuie plus spécifiquement sur la terminologie issue des textes en ancien français, la section 2 décrit le processus d'extraction mis en place. Avant d'extraire la terminologie, il est nécessaire d'utiliser un étiqueteur grammatical. Deux étiqueteurs seront appliqués à notre corpus médiéval : l'étiqueteur de Brill (section 3) et le TreeTagger (section 4). Le système TERVOTIQ qui permet de classer les termes extraits à partir des textes traités par plusieurs étiqueteurs est présenté en section 5. Enfin, la section 6 présente quelques perspectives.

[1] http://album.revues.org/index3032.html

[2] http://weblex.ens-lsh.fr/doc/weblex/outils.html#lml

[3] Pour la justification de ce format, voir l'article de (Heiden et Guillot, 2003).

2. Extraction de la terminologie

2.1. Approche d'extraction de la terminologie

Dans nos travaux, nous utilisons le système EXIT (Roche et al., 2004) pour extraire la terminologie. À partir d'un corpus étiqueté grammaticalement (voir section suivante), le système permet entre autres d'extraire les candidats termes respectant des patrons syntaxiques définis (Nom-Nom, Nom-Préposition-Nom, etc). La section suivante présente les systèmes d'étiquetage grammatical les plus usités.

2.2. Étiquetage grammatical

Cette section présente deux étiqueteurs qui seront utilisés pour traiter des textes en français médiéval : le TreeTagger et l'étiqueteur de Brill.

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 étiquetages corrects. 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).

N'ayant pas de corpus étiquetés manuellement en relation directe avec le corpus spécialisé étudié, nous ne pouvons mettre en oeuvre une phase d'apprentissage supervisé comme dans les travaux de (Stein, 2003). Dans un premier temps, notre approche consiste à construire un lexique adapté à l'ancien français. La méthode mise en place pour construire ce lexique qui sera utilisé par l'étiqueteur de Brill est détaillée dans la section suivante (section 3). Précisons que le TreeTagger utilise son propre lexique adapté à l'ancien français comme nous le montrerons dans la section 4.

3. Utilisation de l'étiqueteur de Brill

3.1. Construction d'un lexique

Afin de réaliser un étiquetage de bonne qualité des textes en ancien français, nous utilisons deux lexiques. Le premier est un lexique en français moderne contenant plus de 440 000

formes fléchies obtenu auprès de l'INaLF (Institut National de la Langue Française[4]). Le second est un lexique en ancien français qui contient un peu plus de 45 000 formes fléchies. Celui-ci est issu des travaux de Mr Douglas C. Walker de l'Université de Calgary (données disponibles à l'URL : http://www.acs.ucalgary.ca/ dcwalker/Dictionary/dict.html). Dans ces lexiques chaque mot est associé à une ou plusieurs étiquettes. Par exemple, le lexique en français moderne possède, pour chaque ligne, la structure suivante : *mot étiquette*$_1$ *étiquette*$_2$ *... étiquette*$_n$. 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. Afin d'améliorer la qualité de l'étiquetage, nous avons fusionné les deux lexiques. Lors de cette fusion, les mots et étiquettes présents dans le lexique AF sont privilégiés par rapport aux mots et étiquettes du lexique FM.

Notons que dans certains cas, nous pouvons avoir un mot AF égal à un mot FM mais avec des étiquettes différentes. Dans le lexique en français moderne, les étiquettes sont associées au mot selon un ordre de probabilité (les étiquettes les plus probables pour un mot précèdent les moins probables). Malheureusement, cette information ne nous est pas donnée par le lexique en ancien français. Par conséquent, dans cette situation, nous positionnons les étiquettes communes au mot AF et au mot FM dans l'ordre donné par le mot FM (ce qui semble plus adapté qu'un ordonnancement des étiquettes de manière aléatoire). Par exemple, le mot AF « **de** Adverbe Preposition » et le mot FM « **de** Preposition Determinant » co-existent. L'occurrence de sortie dans le lexique mixte sera donc « **de** Preposition Adverbe », les étiquettes AF étant toujours privilégiées lors de la construction du lexique mixte.

L'algorithme appliqué pour réaliser la fusion se compose de trois étapes :

1. La première étape consiste à vérifier l'existence de chacune des occurrences du lexique AF dans le lexique FM. Si les mots sont égaux, nous vérifions d'abord la présence d'étiquettes communes entre ces mots. Si la similitude d'étiquettes est avérée, nous les positionnons dans l'ordre du mot FM, sinon nous écrivons dans le lexique mixte l'occurrence du lexique AF.
2. Cette deuxième étape ajoute dans le lexique mixte toutes les entrées du lexique AF qui ne sont pas dans le lexique FM.
3. Puis, la troisième étape complète le lexique mixte par les mots du lexique FM qui ne sont pas dans le lexique AF.

Le lexique AF contenait 2 138 mots qui existaient aussi dans le lexique FM ce qui représente 4,7%. 43 083 entrées du lexique AF étaient quant à elles inconnues du lexique FM, ce qui représente 95.3%. Ainsi, ces entrées qui n'apparaissaient pas dans le lexique FM ont été intégrées dans le lexique mixte. Même si cette fusion des lexiques était utile, nous pouvons regretter le fait que le lexique AF ne représente qu'un neuvième du lexique FM en terme de nombres d'entrées.

3.2. Expérimentations de l'étiquetage de Brill avec le lexique constitué

Deux types d'expérimentations sont mises en oeuvre pour évaluer la qualité du lexique mixte créé. Les premières expérimentations consistent à effectuer une étude quantitative en mesurant le taux de couverture (section 3.2.1). Les expérimentations suivantes s'appuient sur

[4] http://www.inalf.cnrs.fr

une étude qualitative par l'expertise manuelle de l'étiquetage obtenu en utilisant le lexique mixte (section 3.2.2).

3.2.1. Evaluation 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 calcul sont proposés. La première méthode consiste à prendre seulement en compte les mots uniques (aussi appelés *types*). 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.

$$TdC_{type} = \frac{nombred'occurrencesuniques(type)dutexteprésentesdansnotrelexique}{nombred'occurrencesuniques(type)dutexte} \quad (1)$$

Le second calcul du taux de couverture donné par la formule (2) prend en compte tous les mots du corpus (aussi appelés *occurrences*). 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 $TdC_{occurrences}$. Dans le cas où les mots fréquents sont présents dans le lexique, $TdC_{occurrences}$ aura alors une valeur bien supérieure à TdC_{type}. L'objectif ici est d'évaluer si les lexiques utilisés couvrent les mots principaux de l'ancien français.

$$TdC_{occurrences} = \frac{nombred'occurrencesdutexteprésentesdansnotrelexique}{nombred'occurrencesdutexte} \quad (2)$$

L'exemple ci-dessous 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 soumiers **estoit moult bien** cargiés **de** joiaus **et de vaselemente et de deniers**.

À partir de cet exemple, nous pouvons établir les taux de couverture suivants :

$$TdC_{type} = \frac{8}{11} \quad donc \quad TdC_{type} = 0.72 \quad TdC_{occurrences} = \frac{11}{14} \quad donc \quad TdC_{occurrences} = 0.78$$

Le calcul de la couverture en utilisant différents lexiques est donné dans le tableau 1. Ce dernier montre que les lexiques en ancien français et en français moderne couvrent de façon identique le corpus (*TdC* du même ordre pour AF et FM). Ceci ne signifie pas pour autant que les deux lexiques comportent ou identifient les mêmes mots. Par exemple, le $TdC_{occurrences}$ entre le lexique AF et le lexique FM sont similaires (68% versus 70%) alors que le lexique AF ne contient qu'un neuvième du nombre d'entrées du lexique FM.

	Lexiques		
	AF	FM	Mixte
TdC_{type}	35%	37%	53%
$TdC_{occurrences}$	68%	70%	81%

Table 1 : Taux de couverture

Par ailleurs, le tableau 1 montre que le lexique mixte que nous avons construit couvre davantage le corpus comparativement à l'utilisation des lexiques AF ou FM. Ce taux de couverture est d'ailleurs très important avec $TdC_{occurrences}$ (81%) ce qui montre que les mots les plus courants de l'ancien français sont présents dans le lexique mixte constitué.

3.2.2. Évaluation qualitative de l'étiquetage grammatical

Après avoir évalué quantitativement notre méthode de construction du lexique mixte, une évaluation qualitative est présentée dans cette section. Cette évaluation manuelle va être appliquée à des textes étiquetés en utilisant différents lexiques. À partir de l'évaluation manuelle, 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 menée (voir tableau 2).

Lexiques	**Taux d'erreur de l'étiquetage**
Ancien français (AF)	46%
Français moderne (FM)	63%
Mixte	35%

Table 2 : Taux d'erreur de l'étiquetage grammatical

Les résultats du tableau 2 montrent que le taux d'erreur est beaucoup plus faible avec le lexique Mixte comparativement aux lexiques AF et FM.

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.

L'utilisation des différentes connaissances des deux lexiques diminue significativement le taux d'erreur qui reste pourtant assez important (35%). 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 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 ainsi 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. Une phase d'apprentissage supervisé pourrait alors être menée. Cette phase nécessiterait l'utilisation d'un corpus d'une taille conséquente entièrement annoté.

4. Étiquetage avec le TreeTagger

Après avoir appliqué l'étiquetage de Brill dont le lexique a été adapté à l'ancien français, nous utilisons le TreeTagger qui utilise des ressources issues de l'ancien français. Le lexique utilisé

par le TreeTagger provient en effet 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 formes fléchies d'ancien français. 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 (Nancy, INaLF/CNRS)*. Au final, le lexique en ancien français que nous avons utilisé contient un peu plus de 234 000 formes.

Nous avons relevé dans la section précédente que l'utilisation de l'étiqueteur de (Brill, 1994) présentait un taux d'erreur de 35%. Or, le TreeTagger dispose d'un lexique en ancien français conséquent suite, entre autres, à une mise à jour récente (octobre 2006). Nous avons donc appliqué le TreeTagger avec son lexique et ses règles d'étiquetage adaptées à l'ancien français en l'état sur notre corpus du Lancelot. À partir de ce fragment, une analyse des erreurs d'étiquetage a été menée afin de déterminer leurs origines. Il ressort de cette analyse un taux d'erreurs de 21% avec le TreeTagger, ce qui est nettement meilleur qu'avec Brill (35%). Les deux étiqueteurs s'appuient sur des données différentes. Dans le cas du TreeTagger, ces dernières sont parfaitement adaptées à l'ancien français d'où un étiquetage naturellement de meilleure qualité. Ainsi, nos travaux ne concluent pas sur le fait que l'étiqueteur de Brill est moins adapté à l'ancien français car nous n'avons pu entraîner cet étiqueteur à partir de l'ancien français (car nous n'avons pas à notre disposition un corpus conséquent entièrement annoté manuellement).

Les deux étiqueteurs présentés dans les sections précédentes s'appuient sur des approches différentes et utilisent des ressources distinctes. Nous allons utiliser cette situation complexe relative à l'indépendance importante des étiqueteurs comme un avantage pour l'extraction de la terminologie. Ainsi, nous pouvons avoir une confiance élevée pour les termes extraits à partir des textes étiquetés par ces deux approches significativement différentes.

5. Combinaison des étiqueteurs pour l'extraction de la terminologie : le système TERVOTIQ

L'approche présentée dans cette section, appelée TERVOTIQ (TERminologie par VOte selon l'éTIQuetage) s'appuie sur un système de vote (Màrquez et al., 1998 ; Illouz, 1999 ; Sjöbergh, 2003). En effet, la combniaison d'étiqueteurs (Illouz, 1999) ou d'analyseurs syntaxiques (Monceaux, 2002) donne en général des résultats particulièrement intéressants d'où notre objectif de mettre en place un système de vote fondé sur des étiqueteurs indépendants.

5.1. Attribution des indices de fiabilités initiaux

L'objectif de TERVOTIQ est de déterminer la pertinence de l'ensemble des termes extraits qu'ils soient fréquents ou non. La méthode de validation retenue repose sur un protocole spécifique. Ce protocole attribue un indice de fiabilité à chaque candidat terme extrait en utilisant un texte traité avec un étiqueteur donné. L'attribution d'un indice à un étiqueteur a été réalisée suite à l'évaluation par l'experte médiéviste du fragment étiqueté par Brill puis par le TreeTagger (voir sections 3 et 4). Cette évaluation manuelle a montré que le taux d'erreur de Brill était plus élevé que celui du TreeTagger, c'est la raison pour laquelle l'indice

de Brill a été attribué à 1 et celui du TreeTagger à 2. Plus l'indice est élevé, plus nous pouvons attribuer une confiance importante aux candidats termes extraits.

Grâce au système TERVOTIQ, nous pouvons attribuer une note à chacun des candidats termes extraits. À partir des textes étiquetés par les deux systèmes (TreeTagger et Brill), deux listes de candidats termes ont été retournées par EXIT. Il est alors possible de déterminer :

(1) l'indice global des candidats termes communs (qui apparaissent dans les deux listes), (2) l'indice spécifique aux candidats termes appartenant à une seule des deux listes (appelés candidats termes « complémentaires »). Ce système repose donc sur la distinction de ces différents candidats termes.

5.2. Combinaison des indices de fiabilité

Notre système s'appuie sur l'intersection des candidats termes extraits en leur attribuant la somme des indices de fiabilité initiaux. Afin d'illustrer de façon concrète le fonctionnement du système TERVOTIQ, prenons un exemple qui utilise trois étiqueteurs : *EtiqA, EtiqB et EtiqC*. Un indice de fiabilité différent est attribué à chacun des étiqueteurs : *EtiqA* = 1, *EtiqB* = 2, *EtiqC* = 35. Nous avons quatre cas possibles concernant l'intersection des listes de candidats termes et un cas hors intersection :

1. *EtiqA* ∩ *EtiqB* : l'indice de fiabilité vaut : 1 + 2

2. *EtiqB* ∩ *EtiqC* : l'indice de fiabilité vaut : 2 + 3

3. *EtiqA* ∩ *EtiqC* : l'indice de fiabilité vaut : 1 + 3

4. *EtiqA* ∩ *EtiqB* ∩ *EtiqC* : l'indice de fiabilité vaut : 1 + 2 + 3

5. si le candidat terme n'apparaît pas dans une intersection, alors il conserve l'indice de fiabilité initialement associé à l'étiqueteur.

Cet exemple permet d'avoir un indice de fiabilité variant de 1 (fiabilité la plus faible) à 6 (fiabilité la plus élevée). Notons que dans un cas d'égalité des indices, par exemple dans notre cas lorsque l'indice 3 de *EtiqC* est égal aux indices 1+2 de *EtiqA* ∩ *EtiqB*, nous privilégions la somme composée par plusieurs indices (soit : 1 + 2 = 3) par rapport à un seul indice (soit : 3).

La principale différence entre les travaux de (Illouz, 1999) et de (Sjöbergh, 2003) et notre système réside dans le fait que nous ne cherchons pas à mettre en place un système de vote retournant des résultats d'étiquetage de meilleure qualité mais à déterminer les termes les plus pertinents en appliquant différents étiqueteurs. En effet, notre objectif est de fiabiliser l'extraction de la terminologie ; le système de vote que nous proposons dans cet article est donc fondé sur cette tâche.

5.3. Classement des termes avec TERVOTIQ

La démarche décrite dans la section précédente a été appliquée aux listes de candidats termes extraits avec EXIT à partir des corpus étiquetés par Brill puis par le TreeTagger. Le classement des termes s'effectue dans un premier temps selon l'indice de fiabilité établi puis les candidats

[5] Soulignons que ces taux pourraient être déterminés de manière automatique sur la base des résulats obtenus sur les corpus de test.

termes ayant le même indice de fiabilité sont classés selon le nombre d'occurrences décroissant (cf. tableau 3)[6].

À partir de ces listes de candidats termes, il est possible de définir plusieurs mesures d'évaluation. Nous retiendrons la mesure de précision qui permet de définir l'aptitude de notre système à ne retenir que les candidats termes pertinents (appelés « termes »). Dans notre cas, un candidat terme est pertinent si le groupe de mots extrait a un sens par rapport aux patrons syntaxiques proposés. Cette définition est donc très large et mériterait d'être approfondie en mettant en place une sémantique plus précise de la notion de pertinence propre à la thématique de parenté des textes médiévaux étudiés. Ainsi, dans nos travaux, nous ne nous intéressons pas spécifiquement à l'extraction des collocations au sens de (Clas, 1994). En effet, dans ce cas, une collocation est définie comme un groupe de mots ayant un sens global qui est déductible des unités (mots) composant le groupe. Par ailleurs, le sens des mots qui composent la collocation doit être limité (Clas, 1994). Dans nos travaux, nous ne souhaitons pas distinguer ces différentes situations définissant une collocation car tous les groupes de mots respectant des patrons syntaxiques, ce qui représente un autre critère définissant une collocation (Clas, 1994 ; Laurens, 1999), sont susceptibles d'être utiles pour l'expert. Ainsi, dans cet article, nous utiliserons le mot « terme » ou « collocation » mais au sens plus large décrit par (Grossmann et Tutin, 2003).

	Nb occurrence(s)	**Terminologie commune**	**Validation**	**Type de relation**
Indice : 3	5	tel manière	-	Adjectif-Nom
	3	grant proeche	+	Adjectif-Nom
	3	grans proeches	+	Adjectif-Nom
	3	grant biauté	+	Adjectif-Nom
	…			
	2	grant riqueche	+	Adjectif-Nom
	2	boines trieves	+	Adjectif-Nom
	…			
	1	pas talent	-	Nom-Nom
	1	pardurable clarté	+	Adjectif-Nom
	1	hons liges	-	Nom-Adjectif
	…			
Indice : 2	12	roi ban	+	Nom-Nom
	4	grant paour	+	Adjectif-Nom
	3	roi bohort	+	Nom-Nom
	...			
Indice : 1	10	l' abeesse	-	Nom-Nom
	...			

Table 3 : Extrait de candidats termes binaires classés par nombre d'occurrences pour chaque indice.

Précisons que nous ne retiendrons pas la mesure de rappel car ce critère d'évaluation nécessite de connaître l'ensemble des candidats termes pertinents qui peuvent se révéler très nombreux[7].

[6] Si deux termes communs ont un nombre d'occurrences différent selon les étiqueteurs, le nombre d'occurrences de l'étiqueteur le plus fiable est privilégié.

Le tableau 4 présente quelques résultats obtenus à partir de notre corpus. Dans un premier temps, nous avons calculé la précision obtenue pour chaque indice de fiabilité. Un élagage a été appliqué consistant à conserver les candidats termes ayant une fréquence supérieure à un seuil déterminé. Pour les indices 1 et 2, nous avons appliqué un élagage de 2 (les candidats termes n'apparaissant qu'une seule fois ne sont pas considérés). En effet, le nombre de candidats termes était trop important sans appliquer d'élagage pour que l'experte puisse analyser l'ensemble des candidats retournés. Pour l'indice 3, nous avons mené des expérimentations consistant à n'appliquer aucun élagage puis à effectuer un élagage de 2.

Le tableau 4 montre que plus l'indice est important et plus la précision est élevée. Les systèmes de vote uniquement fondés sur l'évaluation de l'étiquetage grammatical ont également montré les résultats de bonne qualité dans le cas de prédictions d'une même étiquette avec plusieurs étiqueteurs (Sjöbergh, 2003). Par ailleurs, ce tableau montre que la précision des termes avec un indice de 3 est plus élevée en appliquant un élagage. Ceci confirme le résultat bien connu fondé sur le fait que les termes les plus fréquents sont globalement les plus pertinents (Roche et Kodratoff, 2006)[8]. Notons que les candidats termes communs qui apparaissent une seule fois étaient initialement mêlés aux nombreux candidats ayant également une occurrence à 1 et qui sont globalement moins pertinents que les candidats fréquents. À l'aide de TERVOTIQ, ces termes communs qui sont rares sont mis en valeur car ils sont présents parmi les candidats termes à indice 3. Enfin, le tableau 4 montre que les candidats termes extraits à partir des corpus traités par l'étiqueteur de Brill qui ne sont pas retrouvés par le TreeTagger sont non pertinents. Cependant, l'utilisation de cet étiqueteur par notre système de vote permet d'améliorer significativement l'approche d'extraction de la terminologie (*i.e.* précision propre à l'indice 3 élevée).

Indice	**Termes binaires**			
	Elagage	Candidats termes extraits	Candidats termes pertinents	Précision
3 (intersection des 2 étiqueteurs)	1	128	95	**74,22%**
3 (intersection des 2 étiqueteurs)	2	19	16	**84,21%**
2 (TreeTagger seulement)	2	17	10	**58,82%**
1 (Brill seulement)	2	30	0	**0%**

Table 4 : Tableau récapitulatif de la précision par indice.

Les expérimentations présentées dans cet article s'appuient sur un sous-ensemble (11 725 mots) du corpus en ancien français. Nous n'avons pu mener les expérimentations à grande échelle avec TERVOTIQ sur l'ensemble du corpus (706 938 mots). En effet, ces expérimentations demandent un travail conséquent d'expertise et d'analyse à partir de la très grande quantité de termes extraits. Par exemple, en nous appuyant sur l'ensemble du corpus étiqueté par le TreeTagger plus de 15 000 termes (sans appliquer d'élagage) ont pu être extraits : 4 662 termes « Adjectif Nom », 1 510 termes « Nom Adjectif », 4 182 termes

[7] Le rappel pourrait être calculé sur un échantillon de corpus représentatif. Cependant, outre le fait que le choix d'un tel échantillon peut se révéler extrêmement complexe à établir, le choix des paramètres est aussi une difficulté majeure. Par exemple, le cas d'un élagage de 2 adapté à notre corpus de test n'est pas nécessairement approprié à un échantillon de corpus, ce qui peut significativement influencer le calcul du rappel. Pour toutes ces raisons, nous avons préféré omettre l'évaluation du rappel et nous appuyer sur la précision.

[8] Précisons cependant que pour certaines tâches spécifiques, les termes fréquents ne sont pas toujours parfaitement adaptés (Roche, 2006).

« Nom Nom », 4 868 termes « Nom Préposition Nom ». L'analyse de ces termes représente une des perspectives importante du travail présenté dans cet article.

6. Conclusion et perspectives

Cet article présente un système de vote d'étiqueteurs grammaticaux pour l'extraction de la terminologie à partir d'un corpus en français médiéval. Les étiqueteurs appliqués utilisent des méthodes mais également des ressources différentes. Pour utiliser l'étiqueteur de Brill, nous avons construit un lexique adapté à l'ancien français. Cependant les règles lexicales et contextuelles de l'étiqueteur de Brill que nous avons appliquées ne sont pas adaptées à l'ancien français. Le TreeTagger que nous avons également utilisé a quant à lui été entraîné à partir d'un corpus en ancien français. Cet étiqueteur utilise donc des règles mais également un lexique adaptés à notre corpus.

Dans nos travaux, nous avons mis en place un système de vote appelé TERVOTIQ qui utilise le résultat de plusieurs étiqueteurs. Notre système qui s'appuie sur un indice de fiabilité montre des résultats intéressants afin de proposer une terminologie pertinente. Ces résultats encourageants pourraient s'expliquer par le fait que les termes communs trouvés par plusieurs étiqueteurs indépendants sont plus pertinents. En effet, un même type de bruit propre à l'étiquetage étant rarement produit par plusieurs étiqueteurs indépendants, le système TERVOTIQ permet un filtrage en mettant en relief des termes plus fiables, c'est-à-dire les termes communs extraits à partir de textes étiquetés de manière différente.

Afin d'augmenter la qualité de l'extraction de la terminologie, il serait intéressant d'améliorer la tâche d'étiquetage grammatical. Pour cela, il serait intéressant d'entraîner l'étiqueteur de Brill sur des données en ancien français afin d'obtenir des règles lexicales et contextuelles adaptées. Par ailleurs, nous pourrions appliquer TERVOTIQ avec d'autres étiqueteurs afin d'étudier la validité à plus grande échelle de l'indice de fiabilité proposé. Enfin, la terminologie extraite pourrait être associée à des méthodes de fouille de données afin d'identifier des liens entre différents concepts (regroupement de termes). Cette analyse permettrait de mettre en exergue des connaissances nouvelles et intéressantes pour les experts.

Références

Brill E. (1994). Some advances in transformation based part of speed tagging. In *AAAI,* Vol.1, pages 722-727.

Cazal E., Serp C., Roche M. and Laurent A. (2007). Extraction de terminologie pour l'ancien français : la quête du graal. In *Actes de l'atelier FDC'07 (Fouille de Données Complexes dans un processus d'extraction des connaissances) à la conférence EGC'2007*, pages 11-20.

Clas A. (1994). Collocations et langues de spécialité. *Meta*, 39(4), 576-580.

Grossmann F. and Tutin A. (2003). Les collocations : analyse et traitement. Amsterdam : Editions De Werelt. Collection : *Travaux et recherches en linguistique appliquée.* Série E : *Lexicologie et lexicographie, ISSN 1572-042X.*

Heiden S. et Guillot C. (2003). Capitalisation des savoirs par le 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.* P. Kunstmann, F. Martineau & D. Forget Eds.

Heiden S. (2004). Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex. In *Actes des 7e Journées internationales d'Analyse Statistique des Données Textuelles*

(JADT'04), Le poids des mots. Vol. 1, pages 577-588, Gérard Purnelle, Cédrick Fairon, Anne Dister, Eds, Presses Universitaires de Louvain, Louvain-la-Neuve, Belgique.

Illouz G. (1999). Méta étiqueteur adaptatif : vers une utilisation pragmatique des ressources linguistiques. In *TALN99, Cargèse*.

Laurens M. (1999). La description des collocations et leur traitement dans les dictionnaires. *Romaneske*, 4.

Màrquez L., Padro L. and Rodriguez H. (1998). Improving tagging accuracy by using voting taggers. In *Proceedings of NLP+IA/TAL+AI'98.* Moncton, New Brunswick, Canada.

Monceaux L. (2002). *Adaptation du niveau d'analyse des interventions dans un dialogue – application à un système de question-réponse*. PhD thesis, Université Paris Sud.

Quinlan J. R. (1986). Induction of decision trees. *Machine Learning*, 1, 81-106.

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).

Roche M. (2006). Acquisition de la terminologie et définition des tâches à effectuer, deux principes indissociables. In *Actes des Journées de Rochebrune (Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels)*, p. 151-161.

Roche M., Heitz T., Matte-Tailliez O. and Kodratoff Y. (2004). EXIT : Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. *Proceedings of JADT'04 (Journées internationales d'Analyse statistique des Données Textuelles)*, p. 946-956.

Roche M. and Kodratoff Y. (2006). Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. In *Proceedings of onToContent'06 workshop (Ontology content and evaluation in Enterprise) - OTM'06,* Springer Verlag, LNCS, p. 1107-1116.

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

Serp C. (2007). Le mariage comme impasse narrative dans le cycle lancelot-graal. In colloque *Le mariage dans la littérature narrative avant 1800*.

Sjöbergh J. (2003). Combining pos-taggers for improved accuracy on swedish text. In *Proceedings of NoDaLiDa (14th Nordic Conference of Computational Linguistics)*.

Stein A. (2003). Part of speech tagging and lemmatisation of old french texts. In http://www.unistuttgart.de/lingrom/stein/forschung/altfranz/afrlemma.pdf.