Analyse syntaxique de l'ancien français : quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage ?

Gaël Guibon, Isabelle Tellier, Sophie Prévost, Mathieu Constant, Kim Gerdes

▶ **To cite this version:**

Gaël Guibon, Isabelle Tellier, Sophie Prévost, Mathieu Constant, Kim Gerdes. Analyse syntaxique de l'ancien français : quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage ?. TALN 22, Jun 2015, Caen, France. . hal-01251006

HAL Id: hal-01251006
https://hal.science/hal-01251006

Submitted on 5 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analyse syntaxique de l'ancien français : quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage ?

Gael Guibon [1] Isabelle Tellier [1,2]
Sophie Prevost [1] Matthieu Constant [3] Kim Gerdes [2,4]

(1) Lattice CNRS
(2) université Paris 3 - Sorbonne Nouvelle
(3) université Paris-Est, LIGM
(4) LPP CNRS
E-mails: gael.guibon@gmail.com, isabelle.tellier@univ-paris3.fr,
sophie.prevost@ens.fr, Matthieu.Constant@u-pem.fr, kim@gerdes.fr

Résumé. L'article présente des résultats d'expériences d'apprentissage automatique pour l'étiquetage morpho-syntaxique et l'analyse syntaxique en dépendance de l'ancien français. Ces expériences ont pour objectif de servir une exploration de corpus pour laquelle le corpus arboré SRCMF sert de données de référence. La nature peu standardisée de la langue qui y est utilisée implique des données d'entraînement hétérogènes et quantitativement limitées. Nous explorons donc diverses stratégies, fondées sur différents critères (variabilité du lexique, forme Vers/Prose des textes, dates des textes), pour constituer des corpus d'entrainement menant aux meilleurs résultats possibles.

Abstract.
Old French parsing : Which language properties have the greatest influence on learning quality ?
This paper presents machine learning experiments for part-of-speech labelling and dependency parsing of Old French. Machine learning methods are used for the purpose of corpus exploration. The SRCMF Treebank is our reference data. The poorly standardised nature of the language used in this corpus implies that training data is heterogenous and quantitatively limited. We explore various strategies, based on different criteria (variability of the lexicon, Verse/Prose form, date of writing) to build training corpora leading to the best possible results.

Mots-clés : étiquetage morpho-syntaxique, analyse en dépendance, ancien français, apprentissage automatique, exploration de corpus.

Keywords: POS labelling, Dependency Parsing, Old French, machine learning, corpus exploration.

1 Introduction

L'ancien français a donné lieu à de nombreux travaux linguistiques, mais il a été jusqu'à présent très peu exploré dans une perspective "TAL". Il existe pourtant depuis peu un corpus arboré permettant cette exploration : le SRCMF (Syntactic Reference Corpus of Medieval French), (Stein & Prévost, 2013). Ce corpus, décrit en détail en section 2, contient des textes de divers domaines (littéraire, historique, religieux...), formes (vers/prose), époques (du 10ème au 13ème siècle) et dialectes (normand, champenois, picard...). La langue de ces textes étant beaucoup moins normalisée que maintenant, le SRCMF présente une variabilité et une hétérogénéité qui n'a pas d'équivalent pour le français contemporain.

Les données étiquetées du corpus SRCMF peuvent être utilisées par des techniques d'apprentissage automatique, pour acquérir un étiqueteur morpho-syntaxique et/ou un analyseur en dépendance (le format adopté pour les analyses syntaxiques dans SRCMF) de l'ancien français. Dans (Guibon *et al.*, 2014) nous avons décrit une première série d'expériences exploitant ce corpus. L'emploi de CRF (Lafferty *et al.*, 2001) pour la couche d'annotation en parties du discours, associé à l'utilisation de Mate (Bohnet, 2010) pour les analyses en dépendance, ont ainsi permis d'améliorer les expériences préliminaires réalisées (notamment avec TreeTagger [1]) par Achim Stein (Stein, 2014). Ces séries d'expériences procédaient

1. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

toutes à des "apprentissages croisés" consistant à apprendre à partir d'un des textes à l'exclusion de tous les autres, et à tester chacun des programmes appris sur chacun des autres textes. L'objectif visé était d'étudier les éventuelles corrélations entre les performances de ces programmes et des notions de proximités linguistiques entre textes. Mais ces "apprentissages croisés" étaient aussi en quelque sorte suggérés par les données elles-mêmes. En effet, du fait de la grande hétérogénéité de la langue d'un texte à un autre, procéder à des expériences par "simples" validations croisées (par texte ou en les mélangeant sans discernement) aurait peu de sens. Dans le contexte de ce corpus, il faut porter une plus grande attention que d'habitude aux relations entre les données d'apprentissage et les données de test.

Dans cet article, nous approfondissons ces premières approches en cherchant toujours à obtenir les meilleurs étiqueteurs et analyseurs syntaxiques possibles pour un texte (ou un ensemble de textes) donné(s). Les méthodes d'apprentissage automatique employées ici sont les mêmes que dans (Guibon *et al.*, 2014), mais il ne s'agit plus d'améliorer les résultats obtenus initialement par Stein puisque d'une expérience à une autre nous faisons varier le mode de constitution des données d'entrainement et de test, rendant toute comparaison impossible. Nous cherchons à caractériser finement les propriétés des corpus d'entrainement qui influencent le plus les résultats en test. L'objectif final serait, pour un texte nouveau dont on connaitrait les propriétés principales (en termes de domaine, forme, date d'écriture...) et qu'on voudrait étiqueter au mieux, de constituer pour lui un corpus d'entrainement "sur mesure" à partir des données de SRCMF.

Nous souhaitons ainsi développer une méthodologie applicable lors de l'exploration de n'importe quel corpus hétérogène, en jouant sur les propriétés ou métadonnées qui caractérisent les textes qu'il contient. L'apprentissage automatique est la base de cette méthodologie d'exploration de corpus. Elle gravite ici autour de quatre propriétés ou métadonnées pertinentes pour les données de SRCMF : l'homogénéité lexicale des textes (section 3), la quantité de données en entrainement (section 4), la forme (vers/prose, section 5) et enfin la date d'écriture (section 6) des documents.

2 Le corpus SRCMF

2.1 Présentation du corpus

SRCMF est un corpus d'ancien français annoté syntaxiquement dans le cadre d'un projet ANR-DFG, dirigé par A. Stein [2] (ILR, U. Stuttgart) et S. Prévost [3](Lattice, CNRS/ENS/Paris3) et associant, outre les laboratoires des porteurs, l'ICAR (CNRS/ENS de Lyon). L'objectif de ce projet était la constitution d'une ressource syntaxiquement annotée pour le français médiéval pouvant être utilisable pour l'entraînement ultérieur d'analyseurs syntaxiques. Les ressources initiales utilisées pour la constitution de ce corpus sont la Base de Français Médiéval (BFM [4]) (Guillot *et al.*, 2007) et le Nouveau Corpus d'Amsterdam (NCA [5]) (Stein *et al.*, 2006). La sélection des textes destinés à intégrer le corpus SRCMF s'est faite à partir de différents critères : le caractère « incontournable » de certains textes (la *Chanson de Roland* par exemple), la fiabilité des éditions, la diversité des textes en termes de date, domaine, forme et dialecte et, enfin, le fait que les textes étaient déjà étiquetés morpho-syntaxiquement. Dans un souci d'équilibre entre les textes, ceux qui comprenaient plus de 40 000 mots ont été échantillonnés (début, milieu, fin). La Table 1 présente les principales caractéristiques des textes de SRCMF.

Parmi les 15 textes (245 000 mots) de SRCMF, 10 (soit 201 465 mots) ont été retenus pour nos expériences. Evoquons tout d'abord quelques spécificités de ces données. L'ancien français est un état de langue qui connaît une forte variation morphologique, y compris dans un même texte. Par exemple, l'adverbe 'ainsi' se rencontre sous les graphies suivantes : *ainsi, ainsin, ainsinc , ainssy...* tandis que l'on recense au moins 17 formes différentes pour le pronom personnel 'je' (*je, gié, jou, gel...*). Il intègre un grand nombre de formes contractées (bien plus que le français moderne), qui résultent de phénomènes d'enclise (prise d'appui accentuel d'un mot sur un mot le précédant : *ne + les > nes*, *je + le > jel*). Sur le plan syntaxique, l'expression du sujet n'est pas obligatoire et l'ordre des mots est assez souple (le sujet peut être postverbal et l'objet nominal préverbal), souplesse favorisée par l'existence d'une déclinaison bi-casuelle héritée du latin, mais en voie d'étiolement dès cette époque. Enfin, on y rencontre beaucoup plus de syntagmes discontinus qu'en français moderne :

(1) *et mes sires Gauvains lor demande coment il l'ont **puis** fet **que** il se partirent de cort (Graal, 1230)*
*Et Messire Gauvain leur demande comment ils ont fait **depuis qu'**ils ont quitté la court*

(2) *Si jurroient li compaignon tel serement come **cil** font **qui en queste doivent entrer***
Les compagons faisaient un serment comme font ceux qui doivent entrer en quête

2. http://www.uni-stuttgart.de/lingrom/stein/
3. http://www.lattice.cnrs.fr/Sophie-Prevost,229
4. http://bfm.ens-lyon.fr/
5. http ://www.uni-stuttgart.de/lingrom/stein/corpus/

Texte	Date	Nb mots	Nb clauses	Forme	Dialecte	Domaine
Vie Saint ***Légier***	Fin 10e s.	1388	192	vers	nd (non défini)	religieux
Vie de Saint ***Alexis***	1050	4804	562	vers	normand	religieux
Chanson de ***Roland***	1100	28 766	3857	vers	normand	littéraire
Lapidaire *en prose*	Milieu 12e s.	4708	468	prose	anglo-normand	didactique
Yvain, Chretien de Troyes	1177-1181	41 305	3788	vers	champenois	littéraire
La ***Conqueste*** *de Constantinople* de Robert de Clari	>1205	33 534	2308	prose	picard	historique
Queste del Saint ***Graal***	1220	40 417	3078	prose	nd	littéraire
Aucassin *et Nicolete*	Fin 12e s.-début 13e s.	9844	1101	vers & prose	picard	littéraire
Miracles de Gautier de **Coinci**	1218-1227	17 360	1422	vers	picard	religieux
Roman de la ***Rose*** de Jean de Meun	1269-1278	19 339	1449	vers	nd	didactique

TABLE 1 – Textes du SRCMF utilisés dans nos expériences

2.2 Enrichissement linguistique du corpus

2.2.1 Etiquetage morpho-syntaxique

Le jeu d'étiquettes morpho-syntaxiques utilisé en annotation comprend 60 valeurs [6], structurées en 2 champs : Catégorie et Type, les "catégories" correspondant aux traditionnelles parties du discours (nom, verbe, adjectif, pronom, ...) et les "types" correspondant à la spécification de ces 'catégories' (verbe conjugué : VERcjg, nom propre : NOMpro,...). Pour les formes contractées (phénomènes d'enclise), il existe des étiquettes complexes, qui associent les valeurs des 2 unités linguistiques contractées en une seule unité graphique. Par exemple : *nel* = contraction de *ne* (adverbe de négation) +*le* (pronom personnel) = ADVneg.PROper. Toute unité graphique a une étiquette morpho-syntaxique.

2.2.2 Annotation syntaxique

L'annotation a été réalisée manuellement, en double aveugle, avec le logiciel NotaBene [7] (Mazziotta, 2010), selon un modèle dépendantiel, inspiré de (Tesnière, 1959) et (Polguère *et al.*, 2009). Le modèle utilisé dans SRCMF hiérarchise d'une part des unités syntaxiques : structures (qui ont une tête verbale, nominale, adjectivale ...), nœuds (regroupés dans des structures dépendancielles) et groupes (pour le traitement des faits de coordination). Le modèle hiérarchise d'autre part des fonctions (sujet, objet, modifieur, ...), qui précisent la relation entre le nœud tête et les structures qui en dépendent. Toute relation dépendancielle est ainsi exprimée par un triplet : (nœud mère, nœud fille, relation dépendancielle).

La structure maximale de l'analyse syntaxique est la "phrase" ("Snt" pour Sentence), définie par la présence d'un verbe fléchi, qui la gouverne, et par le fait qu'elle n'a pas de fonction. La "phrase" correspond, dans la grammaire traditionnelle, à une proposition indépendante ou principale. A l'inverse, les propositions subordonnées ont une fonction vis-à-vis d'un élément de la proposition dont elles dépendent (Objet, Complément, Circonstant, Modifieur attaché d'un nom (c'est le cas des propositions relatives)). Dans la terminologie SRCMF, ce sont des "non-phrases". Dans la suite de cet article, nous utiliserons le terme "clause" pour désigner l'ensemble des structures à tête verbale (verbe fléchi), qu'il s'agisse de "phrases" (principales/indépendantes) ou de "non-phrases" (subordonnées). Par ailleurs il n'y a pas de coordination de "phrases", on découpe au maximum. Dans l'exemple suivant, il y a ainsi 4 phrases :

(3) *Lors entre li preudons en sa chapele ,/ et prent .i livre et une estole / et vient fors ,/ et comence a conjurer l' anemi*
Alors, l'homme entre dans la chapelle et prend un livre et une étole, et ressort et commence à exorciser L'ennemi.
(Graal, folio 188b)

Toutes les structures sont annotées, chaque unité a une étiquette syntaxique. Dans les structures non verbales (par exemple SN), on a nécessairement une tête, et l'on peut avoir des ModA (= modifieurs attachés : déterminants, adjectifs, subordonnées relatives) et/ou un RelC/RelNC (relateur (non) coordonnant : préposition, conjonction de coordination ou de subordination). Parmi les fonctions principales des dépendants du verbe, on distingue des Actants (sujet personnel/impersonnel (SjPer/SjImp), Objet (Obj), Complément (Cmpl), Attribut du sujet (AtSj)...), des Auxiliés (Auxilié ac-

6. http ://bfm.ens-lyon.fr/article.php3 ?id_article=176
7. http://sourceforge.net/projects/notabene/

tif/passif (AuxA/AuxP)) et des Circonstants (Circ). Il n'y a pas dans le modèle de SRCMF de tête fonctionnelle : toute structure est gouvernée par un nœud lexical (verbe, nom, etc.) et les nœuds fonctionnels dépendent des têtes lexicales. Ainsi les phrases principales sont gouvernées par le verbe fléchi, qui gouverne le verbe de la phrase subordonnée, et la conjonction (nœud fonctionnel) qui introduit la subordonnée est gouvernée par le verbe de cette dernière. Les groupes prépositionnels sont quant à eux gouvernés par le nom, la préposition dépendant de ce nom. Au final toutes les structures, de type verbal ou non, et quel que soit leur degré d'enchâssement, dépendent, directement ou non, du verbe de la "phrase" (c'est-à-dire du verbe principal). Il existe dans le modèle des *duplicatas*, c'est-à-dire la copie d'un nœud à laquelle on attache une deuxième relation. Ils sont utilisés pour les pronoms relatifs et pour les formes contractées. Ainsi dans (4), le pronom relatif *qui* est un relateur non-coordonnant (RelNC), son duplicata est un sujet personnel (SjPer), tandis que la forme contractée *del* (de + le) est un RelNC et son duplicata un modifieur attaché (ModA).

(4) *…et si demanderoiz Consoil / del roi qui doit venir*
et vous leur demanderez conseil à propos de la venue du roi (Yvain, 1848-1849)

Le logiciel NotaBene exporte les annotations selon deux formats : CoNLL (utilisé pour nos expériences) et TigerXML (utilisé par le logiciel de requête TigerSearch (Lezius, 2002), qui permet une visualisation en graphes). La Figure 1 présente l'arbre au format TigerXML, selon un graphe plus riche en informations que ceux utilisés pour nos expériences.

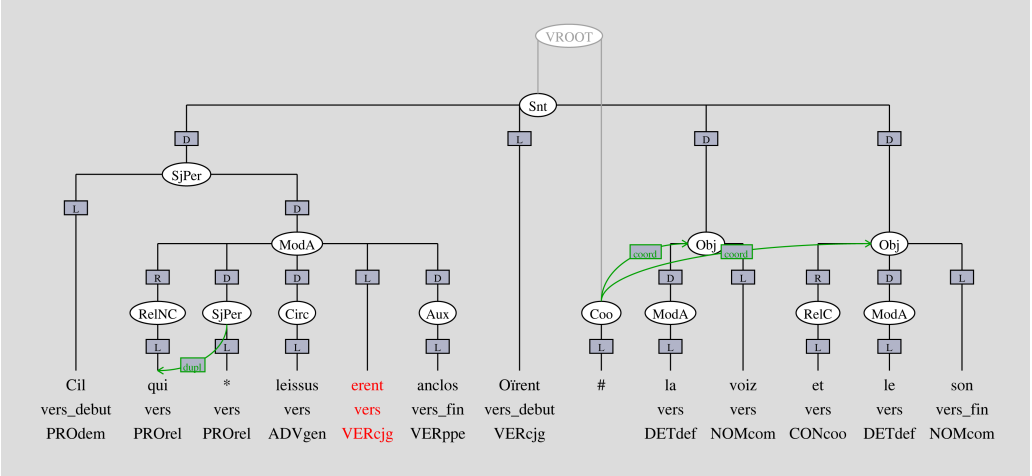

FIGURE 1 – Représentation enrichie d'un arbre syntaxique du SRCMF montrant des cas de duplication

Les graphes correspondent chacun à une phrase pouvant comprendre plusieurs clauses. Il s'agit de clauses verbales gravitant autour d'un verbe conjugué, ce qui engendre des séquences plus courtes que dans des corpus arborés de langue contemporaine tels que le *French TreeBank* (Abeillé *et al.*, 2003), qui sont segmentés en phrases délimitées par une ponctuation. La ponctuation, peu présente dans les fichiers originaux, est absente du SRCMF.

3 Influence du lexique sur l'analyse syntaxique de l'ancien français

3.1 Propriété lexicales

Le corpus présente une grande hétérogénéité lexicale. Afin de mieux la cerner, nous avons mesuré certains paramètres, fournis dans la Table 2. On voit que la taille d'un corpus (en nombre total de mots) n'est pas toujours directement corrélée à sa variabilité lexicale (nombre d'unités distinctes). Le taux de répétition des mots (moyenne, sur des tranches de 1000 mots, du nombre de mots/nombre d'unités) est également très variable d'un texte à un autre. Si on estime que le nombre d'unités distinctes est un bon indicateur de variabilité, c'est *Yvain* qui semble avoir le vocabulaire le plus varié. Cette diversité lexicale, nous l'avons déjà évoqué, est beaucoup plus importante pour l'ancien français que pour le français contemporain, nettement plus standardisé. Même le nom propre *Yvain* peut apparaître dans un même texte sous les formes *Yvain*, *Yvains*, *Yvein* et *Yveins*. La Table 3 montre par ailleurs les taux de vacabulaire commun entre deux textes (nombre de mots dans leur intersection/nombre de mots dans leur union) pour les textes utilisés plus loin.

Corpus	nb unités distinctes	nb mots	nb mots/nb unités	nb répétitions par lemme	taille moy. des clauses
Alexis	1415	4804	3,39	1.98	8
Aucassin	1962	9844	5,02	2.38	9
Coinci	3085	17 360	5,63	2.19	12
Conq.	3424	33 534	9,79	2.9	14
Graal	3874	40 417	10,43	2.56	12
Lapidaire	1197	4708	3,93	2.39	10
Legier	578	1388	2,4	2.08	7
Roland	4304	28 766	6,68	2.00	7
Rose	4097	19339	4,72	2.13	13
Yvain	5040	41305	8,19	2.17	10

TABLE 2 – Quelques propriétés des textes

	Rol.	Graal	Yvain	Conq.
Aucassin	449 /5818 ǀ 7.72%	681 / 5147 ǀ 13.23%	700 / 6289 ǀ 11.13%	665 / 4714 ǀ 14.11%
Roland	/	678 / 7473 ǀ 9.07%	706 /8601 ǀ 8.21%	452 / 7259 ǀ 6.23%
Graal	/	/	1543 / 7340 ǀ 21.0%2	878 / 6411 ǀ 13.70%
Yvain	/	/	/	836 / 7612 ǀ 10.98%

TABLE 3 – Vocabulaire partagé entre deux textes (calcul ǀ %)

Dans cet article, toutes les expériences ont été réalisées selon ce protocole suivant :
– La lemmatisation a été réalisée en utilisant le fichier de paramètres du NCA [8] pour *TreeTagger*, appris par Stein et utilisé lors de ses expériences précédentes (Stein, 2014).
– Les Étiquetages morpho-syntaxiques ont été appris avec *Wapiti* (version 1.4.0) (Lavergne *et al.*, 2010) en utilisant les lemmes prédits précédemment. Les patrons utilisés prennent en compte la vérification du contexte proche, des lemmes avec leur contexte, et de la terminaison des mots, comme dans (Guibon *et al.*, 2014).
– Les analyses en dépendance ont été réalisées avec *Mate* (GraphBased anna-3.61 avec les configurations par défaut) en utilisant les données prédites précédemment.

Nous commençons par reproduire, pour les analyser à l'aune de la variabilité lexicale, quelques résultats d"'apprentissages croisés", dans lesquels un des textes sert d'apprentissage et un autre sert de test. Ces expériences ont été menées d'une part pour l'apprentissage d'un analyseur en parties du discours, d'autre part pour l'acquisition d'un analyseur syntaxique en dépendances utilisant les résultats de l'analyseur en parties du discours.

3.2 Lexique et étiquetage morpho-syntaxique

La prédiction des étiquettes morpho-syntaxiques par "apprentissages croisés" s'est faite sur les cinq plus grands textes disponibles. *Miracles* (Coinci) a été toutefois écarté au profit d'*Aucassin*, car ce dernier présente un mélange intéressant de prose et de vers. Chacun des textes a été découpé en sous-parties d'environ seize mille mots prélevés aléatoirement afin d'équilibrer leurs tailles et donc de limiter autant que possible les biais liés à ces tailles. Les entrainements ont donc toujours utilisé des textes de même taille, chaque modèle appris a été testé sur toutes les autres sous-parties du corpus de test. Les résultats sont présentés dans la Table 4. Chaque valeur est la moyenne des exactitudes des différents modèles.

Dans la Table 4, on voit que pour chaque texte considéré indépendamment en test (en observant le tableau colonne par colonne), le texte permettant d'atteindre en général la meilleure exactitude quand il est utilisé en apprentissage est *Graal*. Ce texte n'est pas le plus riche en nombre d'unités distinctes, mais c'est celui qui présente le plus de répétitions des mêmes mots. Mais *Conquest*, qui présente un taux de répétition (nombre de mots/nombre d'unités) comparable, mène, lui, à des étiqueteurs nettement moins efficaces, ce qui discrédite ce seul indice comme significatif. Plus que la variabilité intrinsèque de chaque texte, c'est sans doute leur proximité deux à deux, lisible dans la Table 3, qui favorise l'apprentissage. Les corrélations sont effectivement assez bonnes, mais parfois irrégulières. Ces premiers résultats contredisent l'hypothèse d'une influence majeure qu'aurait pu avoir un lexique diversifié sur l'étiquetage morpho-syntaxique. Il convient désormais d'aborder son influence sur l'analyse syntaxique afin d'en identifier les spécificités.

8. http://bfm.ens-lyon.fr/article.php3?id_article=324

Train \ Test		Auc.	Rol.	Graal	Yvain	Conq.
Aucassin	Exactitude		80.00	85.76	80.03	87.86
	% Inconnus / Connus		48.19|51.81	29.24|70.76	30.77|69.23	26.99|73.01
	Exactitude I/C		71.43|87.97	72.78|91.07	67.45|85.63	79.31|91.02
Roland	Exactitude	80.48		82.66	78.20	84.13
	% Inconnus / Connus	41.62|58.38		39.90|60.10	40.05|59.95	45.86|54.14
	Exactitude I/C	73.26|85.42		74.45|87.90	73.13|81.42	77.24|89.85
Graal	Exactitude	85.38	80.58		82.70	86.84
	% Inconnus / Connus	30.01|69.99	52.75|47.25		23.11|76.89	32.69|67.31
	Exactitude I/C	73.55|90.45	73.62|88.35		72.78|85.69	78.49|90.88
Yvain	Exactitude	83.13	80.22	89.05		82.11
	% Inconnus / Connus	29.98|70.02	52.00|48.00	19.81|80.19		35.04|64.96
	Exactitude I/C	64.11|91.27	71.20|89.97	74.63 |92.59		66.52|90.52
Conq.	Exactitude	80.48	74.51	79.98	71.04	
	% Inconnus / Connus	31.91|68.09	56.63|43.37	33.43|66.57	37.12|62.88	
	Exactitude I/C	67.42|86.64	67.03|84.28	65.83|87.11	60.78|77.17	

TABLE 4 – Moyenne des différentes expériences d'étiquetages morpho-syntaxiques en "apprentissage croisé"

3.3 Lexique et analyse en dépendance

La diversité du lexique ne semble pas être très influente non plus sur l'analyse syntaxique. Deux méthodes d'évaluation sont utilisées : l'exactitude du gouverneur du mot (*Unlabelled Attachment Score* : UAS) et l'exactitude du gouverneur associé à la bonne fonction syntaxique (*Labelled Attachment Score* : LAS). La Table 5 montre ainsi que plus la méthode d'évaluation est stricte (LAS) plus il apparaît une corrélation entre la diversité lexicale d'un texte et les résultats.

Corpus de test	UAS moyen	LAS moyen	UAS mots inconnus / connus	LAS mots inconnus / connus
Aucassin	75.71 %	61.92 %	67.73% | 78.61%	44.41% | 63.62%
Roland	76.54 %	58.77 %	63.76% | 77.28%	44.10% | 60.36%
Graal	77.62 %	63.94 %	71.53% | 82.74%	49.37% | 69.07%
Yvain	71.14 %	56.03 %	70.53% | 80.75%	46.41% | 65.69%
Conq.	77.67 %	65.67 %	62.98% | 79.08%	41.24% | 66.49%

TABLE 5 – Moyenne des résultats de l'analyse en dépendance par corpus de test

Toutefois l'influence de la diversité lexicale n'est pas isolée et est toujours conjointe avec le nombre de mots des corpus utilisés. La Table 2 montre la nécessité d'étudier ensemble ces deux données. Malgré le fait qu'un corpus d'entrainement ayant un plus grand nombre d'unités différentes devrait être un meilleur candidat à l'entrainement qu'un corpus pauvre en variété, dans les faits le rapport entre le nombre de mots et le nombre d'unités ne permet pas de préjuger d'une bonne analyse en dépendances. Le lexique et le rapport entre le nombre de mots et le nombre d'unités lexicales sur un corpus d'entrainement semble avoir une influence légère sur le score d'évaluation le plus restrictif, le LAS. Au travers des expériences suivantes nous cherchons quelles autres particularités linguistiques du corpus peuvent être les plus influentes.

4 Leave one out

Nous avons d'abord reproduit une expérience de (Stein, 2014) qui consiste à sélectionner un texte faisant office de corpus de test et à lui appliquer le modèle appris conjointement sur les neufs autres textes. Notre expérience, dont les résultats figurent dans la Table 6, diffère cependant de celle de Stein : alors qu'il n'avait pas utilisé le *Roman de la Rose* et avait regroupé ensemble les textes de date inférieure au dixième siècle, nous avons tenu à utiliser également les textes *Legier*, *Alexis* et *Lapidaire*. Autant, dans les expériences précédentes, nous avions privilégié les corpus de taille raisonnable (30 000 mots environ minimum), autant pour celle-ci il nous a paru intéressant d'utiliser aussi les corpus de petite taille. Ces textes, une fois "cumulés", permettent d'obtenir le plus grand corpus d'entrainement possible. C'est cette taille qui compte essentiellement, celle des corpus de test important beaucoup moins. Nous espérons identifier par cette expérience certaines "fractures" (différences significatives) entre un texte particulier et l'ensemble des autres.

XP	Mots inconnus / Mots connus	Exactitude	UAS	LAS
9 sur Alexis	20.05% | 79.95% /	85.91 % 79.01% | 87.12%	81.10 % 72.17% | 83.34%	69.86% 55.56% | 73.44%
9 sur Aucassin	13.92% | 86.08% /	91.21 % 83.58% | 92.45%	86.87 % 74.74% | 88.84%	77.20% 59.56% | 80.06%
9 sur Clari	15.92% | 84.08% /	92.55 % 88.40% | 93.33%	87.12 % 79.52% | 88.63%	78.35% 66.12% | 80.66%
9 sur Coinci	12.96% | 87.04% /	89.72 % 75.51% | 91.84%	79.91 % 66.89% | 81.85%	69.27% 49.64% | 72.20%
9 sur Lapidaire	17.99% | 82.01% /	88.89 % 77.69% | 91.35%	84.88 % 74.26% | 87.21%	75.57% 55.61% | 79.95%
9 sur Legier	59.44% | 40.56% /	66.64 % 52.58% | 76.24%	61.74 % 54.53% | 66.67%	46.04% 33.93% | 54.30%
9 sur Graal	7.17% | 92.83% /	93.58 % 85.51% | 94.19%	89.51 % 80.79% | 90.19%	80.82% 66.06% | 81.97%
9 sur Roland	22.70% | 77.30% /	90.74 % 85.36% | 92.32%	87.91 82.39% | 89.54%	76.23 % 63.82% | 79.88%
9 sur Rose	14.04% | 85.96% /	90.74 % 80.63% | 92.96%	81.56 % 68.89% | 83.64%	70.94% 52.36% | 73.98%
9 sur Yvain	10.55% | 89.45% /	89.61 % 86.11% | 90.02%	84.19 73.76% | 85.42%	74.08% 58.08% | 75.97%

TABLE 6 – Resultats des tests sur un corpus par le modèle appris sur les neuf autres

Les résultats de la Table 6 montrent de fait une fracture visible pour *Legier*, pour lequel aucune des expériences effectuées ne donne de résultats satisfaisants en comparaison avec ceux obtenus avec les autres textes. Avec seulement 66.64 % de taux d'exactitude sur les étiquettes morpho-syntaxiques, il ne faut pas s'attendre à une analyse en dépendances correcte. Là où, pour toutes les autres expériences, la différence entre UAS et LAS avoisine 10 % (réduction également conforme aux résultats cités dans (Stein, 2014)), pour *Legier*, et seulement pour lui, la diminution atteint 15.7 %. On peut imputer cette singularité au fait que ce texte est le plus ancien du corpus. Il se distingue des autres par de grandes différences lexicales et morphologiques, qui influencent directement les résultats de nos expériences.

Cette expérience confirme l'hétérogénéité de l'ancien français. Elle montre qu'un grand corpus d'entrainement n'engendre pas obligatoirement une bonne reconnaissance en test, si les langues diffèrent trop entre les deux. La Table 6 illustre aussi que plus il y a de mots inconnus dans un texte, plus l'écart entre les performances sur les mots connus et inconnus est grand, et plus les résultats sont globalement faibles. Elle invite à chercher d'autres critères, d'autres caractéristiques que la seule taille pour constituer des corpus d'apprentissage "sur mesure". C'est ce que nous proposons dans la suite.

5 Quelles différences d'influence entre vers et prose ?

Nous disposons dans SRCMF de textes pour la plupart en vers, mais également de certains en prose, voire constitués d'un mélange des deux pour *Aucassin*. Cette caractéristique nous a amenés à essayer de quantifier les différences que peut engendrer la forme des textes (vers ou prose) sur l'étiquetage morpho-syntaxique et l'analyse en dépendances.

Corpus	Textes	Nb de mots	Nb unités
Prose	Clari + Graal + Lapidaire	78 904	7271
Vers	Alexis + Coinci + Legier + Roland + Rose + Yvain	113 086	13891
Vers (réduit)	Alexis + Legier + Roland + Yvain	76385	9710
Prose [entrainement]	/	41 910	4320
Vers (réduit) [entrainement]	/	41 907	6840
Prose [test]	/	36 749	4370
Vers (réduit) [test]	/	34 478	4417

TABLE 7 – Corpus de différents types de textes

Le regroupement des textes en vers et en prose conduit à deux corpus de tailles différentes comme on peut le montre la Table 7. Nous avons donc sélectionné un sous-ensemble des textes en Vers permettant un équilibrage des corpus. Cette table nous permet aussi de constater que, même s'ils ont des nombres de mots comparables, les deux corpus Prose/Vers (une fois ce dernier réduit) diffèrent au niveau du lexique. Les textes en vers présentent une plus grande variété lexicale.

5.1 Analyse de l'étiquetage morpho-syntaxique

La Table 8 présente les résultats obtenus en procédant à la division de chaque corpus en deux parties à peu près égales : l'une servant de corpus d'entrainement et l'autre servant de corpus de test pour tous les autres corpus. Il est ainsi possible de comparer l'efficacité de tous les corpus d'entrainement, y compris sur un test de la même forme (Vers ou Prose).

Train \ Test		Prose [test]	Vers réduit [test]
prose	UAS	85.47%	76.33%
	LAS	74.96%	62.96%
	ACC	91.36%	83.61%
	Mots inconnus / Mots connus	16.49% ǀ 83.51%	21.26%ǀ 78.74%
	Lexique différent / commun	57.02% ǀ 42.98%	77.05% ǀ 22.95%
	UAS Mots inconnus / Mots connus	73.76% ǀ 87.78%	65.87% ǀ 79.15%
	LAS Mots inconnus / Mots connus	55.48% ǀ 78.81%	46.37% ǀ 67.44%
	ACC Mots inconnus / Mots connus	77.33% ǀ 94.14%	76.78% ǀ 85.46%
vers réduit	UAS	83.12%	82.79%
	LAS	71.52%	71.40%
	ACC	90.06%	90.78%
	Mots inconnus / Mots connus	18.81% ǀ 81.19%	14.03% ǀ85.97%
	Lexique différent / commun	66,47% ǀ 33,53%	42.52% ǀ 57.48%
	UAS Mots inconnus / Mots connus	73.43% ǀ 85.37%	72.39% ǀ 84.49%
	LAS Mots inconnus / Mots connus	55.45% ǀ 75.24%	55.62% ǀ 73.98%
	ACC Mots inconnus / Mots connus	81.02% ǀ 92.15%	84.13% ǀ 91.86%

TABLE 8 – Resultats des tests entre vers et prose en deux sous-corpus

Nous constatons que le taux d'exactitude reste en moyenne conforme aux expériences faites par apprentissage croisé entre textes. Si l'on prend par exemple les résultats de l'étiquetage morpho-syntaxique appris sur les textes en vers (corpus réduit) et testés sur ceux en prose, les noms communs et les noms propres sont souvent reconnus et bien étiquetés (6.7% d'erreur pour les premiers et 14.6% d'erreur pour les seconds) tandis que ce n'est pas du tout le cas, par exemple, des pronoms possessifs (58% d'erreurs). On observe le même phénomène pour l'expérience inverse avec 10 % d'erreur pour les noms communs, 9% pour les noms propres et 75 % pour les pronoms possessifs.

Surtout, le tableau 8 contredit l'hypothèse selon laquelle la forme d'un texte induirait des analyses syntaxiques différentes. La corrélation entre les résultats ne se fait pas tant en fonction de la forme du texte qu'en fonction du nombre de mots connus et de la diversité lexicale du corpus d'entrainement, comme nous l'évoquions déjà en partie 3. En effet, le corpus d'entrainement en vers comporte davantage de mots communs avec le corpus de test en prose qu'avec le corpus de test en vers. Toutefois cette caractéristique ne se retrouve pas dans le cas du corpus d'entrainement en prose qui offre de meilleurs résultats sur son propre corpus de test (mais ce résultat est toujours lié au nombre de mots connus).

Si l'on regarde le type d'erreurs récurrentes, on voit que ce sont les étiquettes les plus fréquentes qui apparaissent le plus, telle que VERconj (verbe conjugué). Du coup, la fréquence d'erreurs pour les verbes conjugués est basse (2% dans l'expérience du corpus en vers sur le corpus en prose), puisqu'il s'agit d'une des étiquettes attribuées "par défaut". La grande variabilité des formes de l'ancien français joue sûrement un rôle dans ces résultats. Par exemple, les adverbes généraux (ADVgen) se retrouvent régulièrement étiquetés en verbes conjugués, qu'il s'agisse de *Ichi* (ici) ou encore *adont* (donc, alors) comme on le voit dans l'exemple de la Table 9. Ceci s'explique par le fait que *'i'* et *'ont'* sont des désinences verbales. Plus généralement, de nombreux verbes conjugués à la troisième personne du pluriel partagent la terminaison *-ent* avec de nombreux adverbes, engendrant ainsi un mauvais étiquetage. Dans le tableau 9, *COMMENCHE* est étiqueté comme nom propre car la casse de la première lettre d'un mot constitue un des patrons utilisés pour différencier les mots. Retrouver un mot en majuscule étiqueté en tant que nom commun est une erreur régulière, mais, vue la rareté des mots tout en majuscule, enlever ce patron diminuerait les performances globales de l'étiqueteur morpho-syntaxique.

Adont	si	fu	croisiés	li	cuens	Thiebaus	de
VERconj	ADVgen	VERconj	VERppe	DETdef	NOMcom	NOMpro	PRE
Champaigne	ICHI	COMMENCHE	LI	PROLOGUES	DE	COUSTANTINOBLE	
NOMpro	VERconj	NOMpro	DETdef	NOMpro	PRE	NOMpro	

TABLE 9 – Exemples de mauvais étiquetages

5.2 Résultats des analyses en dépendances

Le tableau 8 présente des résultats très bons en comparaison de ceux obtenus lors de la validation inter-textes. Ainsi les scores de l'UAS sont tout particulièrement élevés puisque lors des expériences précédentes, l'UAS maximal obtenu avoisinait les 81%. Plusieurs explications sont possibles : les corpus d'entrainement sont plus grands et couvrent donc davantage de cas différents, ce qui rejoint la question de l'importance du lexique du corpus d'entrainement, sa diversité étant bien plus grande que ce que peut apporter un texte isolé (maximum de 5040 mots différents).

En faisant le rapprochement entre les tableaux 5 et 8, il est possible de conclure sur l'importance d'un corpus d'entrainement au lexique varié, d'autant plus lorsqu'il s'agit de traiter de l'ancien français aux nombreuses variation de formes. Cette variabilité des formes, nous n'avons pu la mesurer que sur le texte de *Yvain*, seul texte pour lequel nous possédions des lemmes vérifiés. Mais, au vu des résultats de l'analyse en dépendances, en particulier le LAS, nous pouvons estimer avoir une plus grande variété de formes dans les textes en vers. Certes, la majorité des textes de l'époque sont en vers, mais en utilisant des corpus de tailles proches nous pouvons observer une diversité lexicale moindre dans les textes en prose. Pour en mesurer l'influence il convient de prendre en considération le lexique commun entre le corpus d'entrainement et le corpus de test, comme cela est montré dans le tableau 8. Ce tableau 8 montre également une corrélation entre taux d'exactitude (accuracy) et UAS obtenue lors de l'analyse en dépendance. Ce n'est toutefois pas toujours le cas pour le LAS, qui semble davantage dépendre de la qualité du lexique plutôt que de la qualité de l'étiquetage morpho-syntaxique.

Train \ Test		Aucassin
prose	UAS	85.72 %
	LAS	75.08%
	ACC	90.89%
	Mots inconnus / Mots connus	18.54% | 81.46%
	Lexique différent / commun	53.00% | 47.00%
	UAS Mots inconnus / Mots connus	74.74% | 88.22%
	LAS Mots inconnus / Mots connus	56.82% | 79.24%
	ACC Mots inconnus / Mots connus	81.04% | 93.13%
vers (réduit)	UAS	83.14%
	LAS	72.48%
	ACC	88.74%
	Mots inconnus / Mots connus	21.26% | 78.74%
	Lexique différent / commun	56.67% | 43.33%
	UAS Mots inconnus / Mots connus	71.62% | 86.25%
	LAS Mots inconnus / Mots connus	56.28% | 76.85%
	ACC Mots inconnus / Mots connus	76.73% | 91.99%

TABLE 10 – Tests des corpus en prose et en vers sur *Aucassin*

Cette expérience sur des corpus de types différents confirme l'influence de la richesse lexicale d'un corpus d'entrainement. En effet, le corpus d'entrainement en vers offre le taux moyen minimal de mots inconnus, sans doute grâce à sa plus grande diversité lexicale (tableau 7). Parmi les trois expériences dépassant le taux de 80% de mots connus, l'entrainement sur le corpus en vers entraine la meilleure reconnaissance syntaxique (LAS) sur les mots inconnus : 2.5% de mieux que lors de l'utilisation du corpus en prose pour l'apprentissage avec test sur d'autres textes en prose, alors que ces derniers comptent une moins grande part de mots inconnus. Nos résultats montrent aussi que lors de l'utilisation d'*Aucassin* (seul texte de SRCMF mélangeant les deux formes prose et vers) en corpus de test, l'entrainement sur la prose donne généralement des résultats meilleurs, corrélés avec le taux plus élevé des mots reconnus. N'ayant pas la quantité exacte de prose et de vers dans *Aucassin*, cette expérience ne peut totalement confirmer les précédentes. Elle montre un score en LAS très proche malgré l'utilisation de corpus d'entrainement distincts avec un taux de mots connus différents.

Il ressort de nos expériences que les textes en vers semblent plus aptes à servir de corpus d'entrainement, tout du moins dans le cadre du corpus SRCMF. Mais les trois textes en prose dont nous disposons couvrent moins d'un siècle de différence, tandis que ceux en vers couvrent presque trois siècles. Cette couverture temporelle différente pour les corpus de chaque forme permet sans doute d'expliquer en partie leur écart conséquent en diversité lexicale. La diversité lexicale si prompte à améliorer les résultats, serait alors liée à l'écart temporel des textes. Ceci nous amène maintenant à explorer l'impact de la période d'écriture sur la capacité à constituer un bon corpus d'entrainement.

6 Une évolution langagière perceptible ?

A l'époque où les textes du SRCMF ont été écrits, peu de personnes savaient lire et écrire. Les échanges se faisaient majoritairement par voie orale, la langue n'était donc que très peu standardisée, d'où les variations d'écriture entre les textes, et sans doute aussi une évolution plus rapide et importante qu'aujourd'hui. C'est particulièrement frappant pour les graphies, certains auteurs utilisant plusieurs orthographes pour un seul mot au sein d'un même texte.

Nous avons tenté de normaliser les formes rencontrées mais avons abandonné cette solution pour deux raisons. La première est linguistique : il est discutable de modifier le corpus traité lorsque nous désirons justement l'explorer tel qu'il est. Une normalisation basée sur nos ressources actuelles devrait être modifiée à chaque ajout de nouvelles données, changeant en même temps le sens des résultats précédents. La seconde raison est pratique : les méthodes simples testées (à base de dictionnaires et de distance d'édition (Myers, 1986)) n'ont pas donné de meilleurs résultats pour les apprentissages.

A la variation lexicale s'ajoute l'existence d'une déclinaison, qui se traduit en particulier par la présence d'un -s désintentiel à la fin de bon nombre de noms masculins singuliers au cas sujet (mais pas au cas oblique (= objet)). Selon la fonction, on peut donc trouver le nom propore *Yvain* (ou *Yvein*) décliné en *Yvains* ((ou *Yveins*), sachant que dès cette époque la déclinaison commence à s'effriter (on a des cas sujets singuliers sans '-s'). De plus la graphie des formes a en partie changé au cours des siècles et, du fait que le SRCMF s'étend sur plusieurs siècles, il nous a paru intéressant d'effectuer une expérience d'apprentissage utilisant des corpus de textes de siècles différents. Cependant, *Vie de Saint Alexis* et *Vie Saint Légier* étant les uniques représentants de la période la plus ancienne, nous ne pouvions regrouper que des textes de deux siècles différents de taille satisfaisante : le 12ème siècle et le 13ème siècle, décrits dans la Table 11.

Corpus \ Infos	Textes	Nb de mots	Nb unités	Taille moy. des clauses
Corpus du 12ème siècle	Lapidaire + Roland + Yvain	74 779	9258	9
Corpus du 13ème siècle	Aucassin + Clari + Coinci + Graal	101 155	9232	12
Corpus 12ème siècle [entrainement]	/	31 488	4901	7
Corpus 13ème siècle [entrainement]	/	50 740	5772	12
Corpus 12ème siècle [test]	/	43 291	5580	9
Corpus 13ème siècle [test]	/	50 415	5226	12

TABLE 11 – Corpus de siècles différents

Pour nos expériences nous avons donc utilisé deux corpus de siècles différents dont la taille et la longueur moyenne des clauses diffèrent, tandis que la diversité lexicale est semblable, ce qui n'était pas le cas en partie 5. Afin de pouvoir comparer leur influence en tant que corpus d'entrainement, nous avons procédé à la même séparation que dans la partie précédente, à savoir la division de chaque corpus en un corpus d'entrainement et un de test mais, contrairement à l'expérience précédente, elle n'induit pas une diversité lexicale très différente entre les différents sous-corpus. Même si le corpus d'entrainement issu des textes du 13ème siècle est apparemment plus "apte" à l'entrainement que celui issu du 12ème siècle de par sa taille et sa richesse lexicale, l'objectif des expériences présentées dans le tableau 12 est de chercher s'il y a une différence notable entre l'utilisation de textes de siècles différents. Plus exactement, il s'agit moins de comparer les corpus d'entrainement que de comparer leurs résultats sur les différents corpus de tests.

Il apparaît ainsi que l'utilisation d'un corpus d'entrainement d'un siècle donné pour un corpus de test du même siècle ne produit pas toujours de meilleurs résultats que son application sur un corpus d'un siècle différent, ce qui peut sembler inattendu. C'est le cas pour l'apprentissage sur le corpus d'entrainement composé de textes du 12ème siècle, qui donne de moins bons résultats sur son propre corpus de test que sur le corpus de test de textes du 13ème siècle. Compte tenu des expériences précédentes, on pourrait penser que la qualité moindre de ces résultats est la conséquence d'une relation

Train \ Test		12ème Sièce [test]	13ème Siècle [test]
12ème Siècle	UAS	66.07%	66.72%
	LAS	50.71%	52.89%
	ACC	71.88%	75.19%
	Mots inconnus / Mots connus	30.88% | 69.12%	32.33% | 67.67%
	Lexique différent / commun	58.13% | 41.87%	62.52% | 37.48%
	UAS Mots inconnus / Mots connus	51.41% | 72.62%	52.00% | 73.75%
	LAS Mots inconnus / Mots connus	33.24% | 58.52%	33.82% | 62.00%
	ACC Mots inconnus / Mots connus	55.83% | 79.05%	57.88% | 83.46%
13ème Siècle	UAS	72.31%	81.15%
	LAS	57.57 %	69.52%
	ACC	78.28%	87.32%
	Mots inconnus / Mots connus	22.52% | 77.48%	17.65% | 82.35%
	Lexique différent / commun	70.17% | 29.83%	52% | 48%
	UAS Mots inconnus / Mots connus	59.44% | 76.06%	68.51% | 83.86%
	LAS Mots inconnus / Mots connus	39.32% | 62.87%	50.11% | 73.69%
	ACC Mots inconnus / Mots connus	62.20% | 82.95%	67.78% | 91.51%

TABLE 12 – Resultats des tests entre deux périodes divisées en deux sous-corpus

moins forte entre les deux textes. Ce n'est cependant pas le cas, le tableau 12 montrant bien un taux de mots connus et de lexique commun toujours plus grand lors de l'application sur un corpus de test du même siècle. Ceci confirme une plus grande relation entre les résultats obtenus et la fréquence des mots déjà connus en entrainement. Cela laisse penser qu'il n'y a donc pas de réelle corrélation entre la proximité temporelle d'un texte et l'efficacité de l'étiquetage morpho-syntaxique et de l'analyse en dépendance. Une autre caractéristique de ces deux corpus pourrait expliquer cet écart au niveau des résultats obtenus : la longueur moyenne des clauses. Nous pouvons nous demander si une longueur moyenne des clauses moindre dans le corpus d'entrainement que dans le corpus de test peut engendrer une difficulté accrue. C'est effectivement le cas en ce qui concerne l'analyse en dépendances, pour laquelle le taux d'erreurs est plus important avec des clauses plus longues dans le corpus de test, mais ceci n'explique pas le taux d'exactitude rencontré.

7 Conclusion

Dans cet article nous avons procédé à l'exploration d'un corpus d'ancien français en utilisant des méthodes d'apprentissage automatique. Diverses expériences ont été menées en vue de quantifier l'impact des caractéristiques de textes de l'ancien français sur les résultats de méthodes d'apprentissage automatique. Nous espérions également tirer de ces exépriences une compréhension meilleure, ou en tout cas différente, de la langue utilisée dans ces textes.

Dans un premier temps, nous avons constaté que la diversité lexicale de la langue n'influe pas clairement sur les résultats et qu'il est nécessaire de l'associer avec le nombre de mots et d'unités présents dans les textes. Dans un second temps, nous avons testé des cas extrêmes d'hétérogénéité (en prenant la totalité des textes hors celui de test en apprentissage), ce qui nous a permis d'isoler le texte le plus ancien et le plus atypique linguistiquement comparé à l'ensemble du SRCMF. Avec les expériences suivantes, nous montrons que la forme (Vers/Prose) d'un texte ne semble pas influencer l'étiquetage morpho-syntaxique et l'analyse en dépendances de l'ancien français. Elles nous ont toutefois permis de constater certaines récurrences au niveau du type d'erreurs et, surtout, de voir l'apparition de ces erreurs dans les deux formes, bien que les textes aient une couverture diachronique différente. Enfin, en essayant de quantifier l'évolution de la langue au cours d'un siècle (évolution qui au final ne s'est pas avérée perceptible), nous observons une diminution de la variété lexicale.

Les résultats du tableau 12 contredisent quelque peu l'influence de la proximité des corpus d'entrainement et de test, observée dans les expériences précédentes. En l'état actuel, nous ne pouvons donc proposer d'explications pleinement satisfaisantes quant aux causes exactes des résultats de l'analyse en dépendances de l'ancien français. Il convient donc d'envisager des expériences complémentaires, afin d'évaluer l'influence possible du domaine et du dialecte des textes.

De manière générale, cet article illustre que l'apprentissage automatique de différents niveaux d'analyse de la langue est une méthodologie puissante pour explorer les propriétés d'un corpus hétérogène. Les stratégies employées dans nos expériences diffèrent uniquement par le choix des corpus servant en entrainement et en test. Les résultats obtenus, de par

leur grande variabilité, donnent des indices précieux pour caractériser les propriétés les plus discriminantes qui distinguent un texte (ou un ensemble de textes) d'un autre (ou d'un ensemble d'autres).

La méthodologie proposée ici est de fait applicable à d'autres corpus puisqu'elle permet d'explorer en isolant les caractéristiques d'un texte de façon ordonnée et hiérarchique à partir du lexique, des étiquettes morpho-syntaxiques, et des structures syntaxiques. Nos expériences suggèrent ainsi que, face à un nouveau texte de l'ancien français, il conviendrait de suivre la méthode suivante :

– Dans un premier temps procéder à une approche générale en effectuant une analyse lexicale complète (lexique et rapport entre nombre d'unité et de mots) et en comparant ce texte aux autres par l'expérience "leave one out" (partie 4).
– Dans un second temps y appliquer une analyse guidée par des caractéristiques linguistiques ou des métadonnées connues du corpus : la date de parution, forme du texte, le dialecte utilisé, etc.

Savoir, par exemple, s'il est préférable de disposer d'un très grand nombre de données annotées ou, au contraire, de données annotées moins nombreuses mais plus proches des données cibles est une question commune à de nombreux domaines (Rafrafi *et al.*, 2013; Tellier *et al.*, 2013). Avec cet article, nous esquissons une méthode générale qui pourra trouver des applications dans d'autres contextes faisant intervenir des corpus hétérogèes

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In *Treebanks*, p. 165–187. Springer.

BOHNET B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *The 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China.

GUIBON G., TELLIER I., CONSTANT M., PRÉVOST S. & GERDES K. (2014). Parsing poorly standardized language dependency on old french. In *13th Treebank and Language Theory (TLT)*.

GUILLOT C., LAVRENTIEV A. & MARCHELLO-NIZIA C. (2007). La base de français médiéval (bfm) : états et perspectives. *Le nouveau corpus d'Amsterdam : actes de l'atelier de Lauterbad, 23-26 février 2006*, p. 143–152.

LAFFERTY J., MCCALLUM A. & PEREIRA F. C. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, p. 282–289, Seattle, Washington.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

LEZIUS W. (2002). Tigersearch ein suchwerkzeug fr baumbanken. In *Proceedings of the 6. Konferenz zur Verarbeitung naturlicher Sprache (6th Conference on Natural Language Processing, KONVENS 2002), Saarbrucken, Germany*.

MAZZIOTTA N. (2010). Logiciel notabene pour l'annotation linguistique. annotations et conceptualisations multiples. In *Recherches qualitatives. Hors-série Les actes*, volume 9.

MYERS E. (1986). An o(nd) difference algorithm and its variations. *Algorithmica*, p. 251–266.

POLGUÈRE A. *et al.* (2009). *Dependency in linguistic description*, volume 111. John Benjamins Publishing.

RAFRAFI A., GUIGUE V. & GALLINARI P. (2013). Classification de sentiments multi-domaines en contexte hétérogène et passage à l'echelle. In *COnférence en Recherche d'Information et Applications (CORIA)*.

STEIN A. (2014). Parsing heterogeneous corpora with a rich dependency grammar. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland : European Language Resources Association (ELRA).

STEIN A. *et al.* (2006). Nouveau corpus d'amsterdam. corpus informatique de textes littéraires d'ancien français (ca 1150-1350), établi par anthonij dees (amsterdam 1987), remanié par achim stein, pierre kunstmann et martin-d. gleßgen.

STEIN A. & PRÉVOST S. (2013). Syntactic annotation of medieval texts : the syntactic reference corpus of medieval french (srcmf). In T. NARR, Ed., *New Methods in Historical Corpus Linguistics*. Corpus Linguistics and International Perspectives on Language, CLIP Vol. 3, P. Bennett, M. Durrell, S. Scheible and R. Whitt (eds).

TELLIER I., DUPONT Y., ESHKOL I. & WANG I. (2013). Adapt a text-oriented chunker for oral data : How much manual effort is necessary ? In *The 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'2013)*.

TESNIÈRE L. (1959). *Eléments de syntaxe structurale*. Librairie C. Klincksieck.