Développement d'un lexique morphologique et syntaxique de l'ancien français

Benoît Sagot

Développement d'un lexique morphologique et syntaxique de l'ancien français

Benoît Sagot
Inria, 2 rue Simone Iff, 75012 Paris, France
benoit.sagot@inria.fr

RÉSUMÉ

Nous décrivons dans cet article notre travail de développement d'un lexique morphologique et syntaxique à grande échelle de l'ancien français pour le traitement automatique des langues. Nous nous sommes appuyés sur des ressources dictionnairiques et lexicales dans lesquelles l'extraction d'informations structurées et exploitables a nécessité des développements spécifiques. De plus, la mise en correspondance d'informations provenant de ces différentes sources a soulevé des difficultés. Nous donnons quelques indications quantitatives sur le lexique obtenu, et discutons de sa fiabilité dans sa version actuelle et des perspectives d'amélioration permises par l'existence d'une première version, notamment au travers de l'analyse automatique de données textuelles.

ABSTRACT

Development of a morphological and syntactic lexicon of Old French.

In this paper we describe our work on the development of a large-scale morphological and syntactic lexicon of Old French for natural language processing. We rely on dictionary and lexical resources, from which the extraction of structured and exploitable information required specific developments. In addition, matching information from these different sources posed difficulties. We provide quantitative information on the resulting lexicon, and discuss its reliability in its current version and the prospects for improvement allowed by the existence of a first version, in particular through the automatic analysis of textual data.

MOTS-CLÉS : Lexique morphologique, Lexique syntaxique, Ancien français.

KEYWORDS: Morphological lexicon, Syntactic lexicon, Old French.

1 Introduction

L'ancien français regroupe l'ensemble des variétés romanes qualifiées de langues d'oïl, qui se sont développées au nord de la France, au sud de la Belgique et dans les îles Anglo-Normandes [1], telles qu'elles étaient parlées du VIIIe au milieu du XIVe siècle environ. L'ancien français se distingue notamment du moyen français, qui lui fait suite, par la présence de déclinaisons nominales. Ancien puis moyen français peuvent être vus comme les ancêtres successifs du français contemporain.

Les deux principales bases de données textuelles, étiquetées semi-automatiquement en parties du discours et en lemmes, sont la Base de Français Médiéval, ci-après BFM (Guillot *et al.*, 2017) [2],

1. Et jusqu'en Angleterre, si l'on tient compte par exemple des lais de Marie de France.
2. http://bfm.ens-lyon.fr

et le Nouveau Corpus d'Amsterdam, ci-après NCA (Stein & al., 2008). Ces bases contiennent respectivement plus de 4 millions et plus de 3 millions de mots. Les deux principaux corpus arborés de l'ancien français sont le Syntactic Reference Corpus of Medieval French, ou SRCMF (Stein & Prévost, 2013) et la partie couvrant l'ancien français au sein du corpus MCVF (Martineau, 2008). Ces deux corpus, dont seul le premier est librement téléchargeable, ne sont pas annotés selon le même guide d'annotation. Enfin, une partie du SRCMF a fait l'objet d'un travail de conversion vers le modèle de l'initiative *Universal Dependencies* (UD) [3]. Toutes ces ressources rassemblent des textes variés, tant sur le plan du style (prose, vers), du genre (littéraire, religieux, historique, didactique), de l'époque (du Xe au XIIIe siècle) que de la géographie dialectale. Toutefois, certains biais sont inévitables, qui affectent en particulier les études linguistiques quantitatives. Ainsi, faute de textes en prose, les premiers siècles de la période ne peuvent être couverts que par des textes en vers.

La disponibilité de ces corpus a permis le développement d'études linguistiques quantitatives et d'outils de TAL. Des expériences d'annotation morphosyntaxiques ont notamment été réalisées par Stein (2014) avec TreeTagger, suivies par celles de Guibon *et al.* (2014, 2015) avec des champs aléatoires conditionnels. Ces expériences ont été toutes deux complétées par une annotation syntaxique en dépendances à l'aide de l'analyseur Mate (Bohnet, 2010) entraîné sur le SRCMF.

Toutefois, le développement de ressources lexicales pour l'ancien français destinées au traitement automatique des langues n'est pas aussi avancé. Pour l'ancien français, on ne peut guère citer que FROLEX [4], lexique librement disponible développé dans le cadre du projet PaLaFra et que ses auteurs définissent comme un lexique morphologique du français du IXe au XVe siècle — mais nous reviendrons sur l'applicabilité en l'espèce de la notion de « lexique morphologique ». Il a été construit automatiquement à partir de textes annotés extraits de la BFM et du NCA, mais également à partir du Dictionnaire de Moyen Français (DMF) [5, 6]. D'autres ressources existent, de natures différentes, notamment dictionnairiques. Nous y reviendrons à la prochaine section. Aucune de ces ressources ne constitue véritablement un lexique morphologique, pas plus qu'un lexique syntaxique. Un tel lexique est pourtant indispensable au développement de certains analyseurs syntaxique. Il permettrait l'amélioration d'outils comme les étiqueteurs morphosyntaxiques, et ouvrirait de nouvelles perspectives en linguistique quantitative, y compris diachronique.

C'est au développement d'un tel lexique que cet article est consacré. Ce lexique, nommé OFrLex, est donc un lexique morphologique et syntaxique de l'ancien français [7], complété par des liens vers les ressources de départ et des gloses pour certaines entrées. Les difficultés principales ont été de trois ordres : (1) la difficulté de transformer des ressources faiblement structurées en données structurées, (2) la non-cohérence des façons dont sont représentées les informations lexicales, rendant délicate la mise en correspondances entre entrées traitant d'un même lexème, et (3) la construction d'informations précises (classes morphologiques, cadres de sous-catégorisation) à partir de ressources ne contenant ces informations que de façon partielle et sous-spécifiée, voire ne les contenant pas du tout. Il en a résulté le développement d'heuristiques et d'outils dédiés et un effort manuel important. Cela fait d'OFrLex un lexique développé ni de façon automatique ni de façon entièrement manuelle.

3. Ce travail de conversion et leurs auteurs sont détaillés sur le site GitHub du corpus converti.

4. `https://github.com/sheiden/Medieval-French-Language-Toolkit`

5. `http://www.atilf.fr/dmf/`

6. Les lexiques LGeRM, développés à l'ATILF, existent sous deux variantes, l'une dite « médiévale » couvre la période 1300-1550, l'autre couvre les XVIe et XVIIe siècles. Ils ne couvrent donc pas l'ancien français.

7. Comme nous le verrons, le volet syntaxique est dans ce travail restreint au lexique verbal. La valence nominale et adjectivale, en particulier, n'est pas encore décrite. C'est naturellement une première étape, mais elle est à la fois nécessaire et, dans un premier temps, suffisante pour le développement d'analyseurs syntaxiques s'appuyant sur des grammaires lexicalisées.

Forme	Fréquence		Étiquette d'origine			Étiquette CATTEXT étendue		Lemme	Source
	BFM	DMF	AFRLEX	BFM	DMF	conv. 1	conv. 2		du lemme
abassera	2	0		*no pos*		*no pos*	OUT	*no lemma*	BFM
abasseur	0	0	NOM		subst. masc.	NOMcom	NOMcom	abasseur	DMF
abasseure	0	0			verbe		VER	abasseurer	DMF
gaiement	0	9			adv.		ADV	gaiement	DMF
gaiement	1	0		ADVgen		ADVgen	APD	*no lemma*	BFM

TABLE 1 – Quelques exemples d'entrées de FROLEX

Nous présentons tout d'abord les ressources dont nous sommes partis pour développer ce lexique et la façon dont nous en avons extrait des entrées lexicales structurées. Nous décrivons ensuite comment nous en avons dérivé OFrLex. Nous terminons par des données quantitatives sur OFrLex suivies d'une discussion sur sa fiabilité et les étapes futures de son développement et de son utilisation.

2 Extraction d'informations à partir de sources hétérogènes

Comme indiqué ci-dessus, **FROLEX** est une compilation de ressources provenant de la BFM, du NCA et du DMF. Il est constitué de plus d'un million d'entrées extensionnelles, c'est-à-dire d'entrées correspondant chacune à une graphie particulière d'une forme fléchie donnée. Les informations disponibles pour chaque entrée varient d'une entrée à l'autre, en fonction de sa source. Les parties du discours ont toutes été converties, parfois de façon sous-spécifiée, dans le modèle CATTEX, étendu par des indications de genre et de nombre, lorsque pertinent. Quelques exemples d'entrées sont montrées dans la table 1, où l'on constate que certaines entrées sont bruitées. De plus, la variété des sources induit des incohérences dans les conventions de lemmatisation, lorsque toutefois un lemme est fourni. Enfin, l'utilisation du DMF comme source a pour conséquence que ce lexique mélange des entrées relevant de l'ancien français et des entrées relevant du moyen français. Il ne s'agit donc pas d'un lexique morphologique à proprement parler : les lemmes présents dans la ressource n'y sont pas représentés par toutes leurs formes fléchies, et certaines formes ne sont pas associées à des informations morphologiques au-delà de la seule étiquette CATTEX.

Pour l'ancien français, le **Wiktionary (anglophone)** contient environ 6500 entrées lexicales intentionnelles (une entrée correspond à un lexème), ainsi que des descriptions formalisées de classes flexionnelles. Par exemple, l'entrée pour *mengier*[8] fournit des formes alternatives (comme *mangier*), une étymologie, une glose en anglais, et les informations nécessaires pour définir sa flexion. Le processus d'extraction que nous avons utilisé est inspiré par celui décrit dans (Sagot, 2014). Nous avons tout d'abord converti le Wiktionary brut, au format wiki, en un fichier XML structuré. Nous avons ensuite extrait des entrées morphologiques complètes à partir de ce fichier XML : chaque entrée est constituée d'une forme de citation, d'un identifiant de classe flexionnelle et de la liste des radicaux ou formes irréguliers le cas échéant. Nous avons alors développé manuellement dans le formalisme Alexina~~FATRSLI~~ (Sagot & Walther, 2013) une grammaire morphologique qui décrit formellement les plus importantes des classes flexionnelles utilisées par Wiktionary.

Le ***Altfranzösisches Wörterbuch*** **de Tobler et Lommatzsch, ci-après TL**, est le dictionnaire d'ancien français de référence. Ses articles sont rédigés en allemand. Nous l'avons utilisé sous deux

8. https://en.wiktionary.org/wiki/mengier#Old_French

Lemma *Lemme*	Haupt-eintrag *Entrée princ.*	Wortart *Catégorie*	Var. *Var.*	Werk *Source*	Band *Volume*	Spalte *Page*	Zeile *Ligne*	IstVar. *Est une var.*
aatir		v.	ahatir	tl	1	31	37	0
aatir	aaatir	v.			1	25	32	1
aatir	atir	v.			1	640	52	1
aatise		s.f.		tl	1	33	34	0
aatison		s.f.		tl	1	33	37	0

TABLE 2 – Quelques exemples partiels extraits de l'index des entrées du Tobler-Lommatzsch

ealemlne s. f. s. chalemine. calemon $. m. [Name eines Vogels : s. A. Delboulle, Rom. XXXI 366 ; A. Thomas, eb. XXXVI 25 260.] calende s. /. s. chalende. calendre s. /. s. chalendre. ealer (nfr. caler) vb. [REW 1487 cafare ; Godefroy VIII 30 (Compl.) 412a] trans. (Segel) niederlassen, streichen : Therfés s'escrie : Cale, cale ! Mes tuit li	calemine s. f., s. chalemine. calemon s. m. [Name eines Vogels : s. A. Delboulle, Rom. XXXI 366 ; A. Thomas, eb. XXXVI 25 260.] calende s. f., s. chalende. calendre s. f., s. chalendre. caler (nfz. caler) vb. [REW 1487 cafare ; Godefroy VIII (Compl.) 412a] trans. (Segel) niederlassen, streichen : Therfés s'escrie : Cale, cale ! Mes tuit li

calemine	NOMcom.f	s.f.	
calemon	NOMcom.m	s.m.	Name eines Vogels
calende	NOMcom.f	s.f.	
calendre	NOMcom.f	s.f.	
caler	VER	vb.	[trans.] (Segel) niederlassen, streichen

TABLE 3 – Extrait du Tobler-Lommatzsch OCRisé avant (gauche) et après (droite) correction partielle, et entrées structurées extraites (bas)

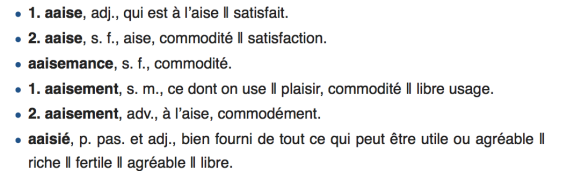
formes, toutes deux produites et distribuées par Achim Stein [9] :

— Une liste des lemmes, constituée en partie manuellement et dans laquelle nous avons trouvé très peu d'erreurs, complétée par un index des formes du *Dictionnaire* de Godefroy pour la fin de l'alphabet [10] (la colonne « Werk » indique la source de chaque ligne : « tl » pour le TL et « g » pour le Dictionnaire de Godefroy). Quelques entrées (simplifiées) de cet index sont fournis à la table 2. On notera que la liste des lemmes distingue les entrées principales (« Haupteintrag ») et les entrées secondaires, ou variantes (généralement des variantes graphiques) : toute entrée secondaire est associée à son entrée principale, et les références (page et ligne) sont données pour l'entrée principale et pour l'entrée secondaire.

— Une version complète OCRisée, qui contient un très grand nombre d'erreurs d'OCR. Nous avons donc corrigé partiellement cette version, de façon manuelle mais systématique, en mettant l'accent sur les parties cruciales des entrées, telles que le type de mot (cf. table 3). Ce travail a été réalisé en alternance avec le développement d'un extracteur d'entrées structurées à partir de la version corrigée du TL OCRisée : cet extracteur effectuant de nombreuses vérifications formelles, il refuse de traiter une entrée dans laquelle il identifie des erreurs. Le résultat de l'extraction, qui inclut des catégories CATTEX, est illustré en bas de la table 3.

Nous avons également utilisé le ***Lexique de l'ancien français*** **de Godefroy**, dans une version publiée sur Wikisource [11], construite au moyen d'un OCR de très bonne qualité et partiellement corrigée

9. https://www.ling.uni-stuttgart.de/institut/ilr/toblerlommatzsch/downloads.htm
10. Ce dictionnaire n'est pas la même ressource que le *Lexique* décrit plus bas.
11. https://fr.wikisource.org/wiki/Lexique_de_l'ancien_français



- **1. aaise**, adj., qui est à l'aise ‖ satisfait.
- **2. aaise**, s. f., aise, commodité ‖ satisfaction.
- **aaisemance**, s. f., commodité.
- **1. aaisement**, s. m., ce dont on use ‖ plaisir, commodité ‖ libre usage.
- **2. aaisement**, adv., à l'aise, commodément.
- **aaisié**, p. pas. et adj., bien fourni de tout ce qui peut être utile ou agréable ‖ riche ‖ fertile ‖ agréable ‖ libre.

FIGURE 1 – Extrait du *Lexique* de Godefroy dans sa version Wikisource

(cf. table 1). Cette ressource est très couvrante mais connue pour contenir des mots (et sens) fantômes. Nous avons donc filtré au moyen de la *Base des mots fantômes [du Godefroy]* [12]. De plus, il couvre jusqu'au XVe siècle, débordant ainsi sur le moyen français. Le caractère structurée du *Lexique* nous a permis d'extraire facilement des entrées structurées combinant une forme de citation, une catégorie CATTEX (complétée le cas échéant d'une information de genre), une définition et l'indication du volume et de la page correspondante.

Enfin, le **Dictionnaire Électronique de Chrétien de Troyes (DÉCT)** est un dictionnaire complet de cet écrivain du XIIe siècle distribué par le CNRTL au format PDF. Nous l'avons converti en format texte, et en avons extrait de façon semi-automatique des entrées structurées à l'aide de règles simples. L'un des intérêts du DÉCT est qu'il relie explicitement ses entrées à des entrées d'autres dictionnaires, dont le TL et le *Dictionnaire* de Godefroy (dont les lemmes sont plus ou moins les mêmes que dans son *Lexique*). Il fournit également les graphies des formes fléchies attestées de chaque entrée.

3 Combinaison des sources et création d'OFrLex

Nous avons tout d'abord relié entre elles les entrées structurées extraites des différentes ressources de la façon suivante, en utilisant comme formes de citation celles du TL. Les entrées du DÉCT pointent vers des entrées du TL, avec très peu d'erreurs. Pour les autres ressources, les lemmes ayant plusieurs entrées conduisent à de multiples correspondances possibles entre entrées issues du Godefroy, du TL et du DÉCT. De plus, pour certains lemmes qui n'ont qu'un lemme dans chaque ressource, ces entrées ne correspondent pas : il faut créer plusieurs entrées et non pas les fusionner. Nous avons donc procédé à une désambiguïsation manuelle, en nous appuyant pour cela sur les définitions contenues dans les différentes ressources. Les informations morphologiques sont extraites des entrées issues du Wiktionary, ou rajoutées automatiquement (lorsque la forme de citation rend cela possible) ou manuellement. Des variantes de formes sont associées à l'entrée grâce aux informations extraites de FROLEX. Le résultat de ce travail est un lexique morphologique où une entrée, qui correspond à un lexème (et non seulement à un lemme) est reliée aux entrées dans les ressources de départ et complétée par les informations que nous y avons extraites, notamment des définitions et gloses, ainsi que des informations sur les variantes (une entrée de type « variante » extraite du TL est associée à son entrée principale, laquelle liste ses variantes). Une catégorie UD, ou UPOS, est également ajoutée automatiquement sur la base des catégories extraites à partir des ressources de départ.

Nous avons alors cherché à compléter les entrées verbales de ce lexique par une couche syntaxique, en suivant les mêmes conventions et critères que ceux du lexique Alexina du français contemporain, le Lefff (Sagot, 2010). Pour cela, nous avons attribué à chaque entrée verbale les informations syntaxiques (valence…) d'un verbe du Lefff susceptible d'être syntaxiquement similaire, ce qui

12. http://stella.atilf.fr/scripts/fantomes.exe

afiner$_1$ v-er 100;Lemma;v;<Suj:cln|sn>;upos=VERB,cat=v;%actif
<link src="TL" loc="TL:1:189:5+1:1224:51" entry="afiner1" ms="v." def="[intr.] enden || [mit pers. obj.] jem. den Garaus machen || [trans. mit sâchl obj.] beenden, zu Ende führen"/> <syntinfosource via="tldef" synttype="I"/>

afiner$_2$ v-er 100;Lemma;v;<Suj:cln|sn,Obj:(cla|sn)>;upos=VERB,cat=v;%actif,%passif
<link src="TL" loc="TL:1:189:47+1:1224:52" entry="afiner2" ms="v." def="[trans.] läutern"/>
<syntinfosource via="tldef" synttype="T"/><hasvariant lemma="effiner" id="1" cat="VER"/>

effiner v-er 100;Lemma;v;<Suj:cln|sn,Obj:(cla|sn)>;upos=VERB,cat=v;%actif,%passif
<link src="TL" loc="TL:1:189:47" entry="afiner2" ms="v." def="[trans.] läutern"/>
<syntinfosource via="tldef" synttype="T"/><variantof lemma="afiner" id="2" cat="VER"/>

effiner, verbe du premier groupe à la flexion régulière
Verbe transitif passivable à sujet nominal ou clitique et à objet direct facultatif nominal ou clitique
Variante de afiner$_2$
Entrée correspondante dans le Tobler-Lommatzsch: afiner2 (1:189:47) '[trans.] läutern'
La valence transitive a été inférée à partir de la glose ci-dessus fournie par le Tobler-Lommatzsch

TABLE 4 – Exemples d'entrées dans OFrLex, suivies d'une version plus explicite de la dernière entrée

UPOS	ADJ	ADP	ADV	CCONJ	DET	INTJ	NOUN	PRON	PROPN	PUNCT	SCONJ	VERB
#entrées	7895	286	1848	37	296	205	44084	517	1948	19	53	16817
#lemmes dist.	7740	283	1804	37	259	174	41191	411	1934	17	50	16152

TABLE 5 – Informations quantitatives sur les entrées d'OFrLex

suppose de lier chaque verbe d'OFrLex à un verbe du L*efff*. Pour cela, pour chaque entrée verbale d'OFrLex, nous avons utilisé, par ordre de priorité décroissant, l'une des informations suivantes :

— Une « pseudo-glose » rajoutée manuellement, choisie exprès pour ses propriétés syntaxiques supposées identiques (ou similaires) à celles de l'entrée d'OFrLex ;
— Une glose en français contemporain issue d'une des ressources de départ ou ajoutée à la main ;
— La définition issue du Godefroy ou du DÉCT dès lors qu'elle est formée d'un seul mot ;
— Un descendant en français contemporain, extrait du Wiktionary ou rajouté manuellement ;
— Une entrée du L*efff* dont la forme de citation est identique à celle d'OFrLex.

Si aucune de ces stratégies n'est applicable, nous avons utilisé les indications de valence extraites des définitions du Godefroy, du TL ou du DÉCT, qui contiennent souvent des étiquettes tels que « trans. », « I » (pour « *intransitiv* » dans le TL) ou « refl. » (sous de multiples variantes). En l'absence de telles informations, nous avons considéré par défaut l'entrée comme transitive simple. On notera que les relations entre variantes sont prises en compte pour récupérer la « meilleure » information possible.

La table 4 contient trois entrées verbales, dont la troisième est une variante de la seconde. Elles sont en deux parties : tout d'abord des entrées Alexina classiques (forme de citation, classe flexionnelle, informations syntaxiques), puis, dans la partie « commentaires » de l'entrée (après le signe #), des éléments XML encodant les informations complémentaires. Pour interpréter les informations syntaxiques on pourra se référer à (Sagot, 2010).

4 Éléments d'évaluation

Des informations quantitatives sur les entrées d'OFrLex, qui correspondent à des lexèmes, sont fournies à la table 5 par catégorie UD. Alexina permet de produire automatiquement à partir de ces entrées intentionnelles près d'un million entrées extensionnelles décrivant chaque (variante de) chaque forme fléchie de chaque lexème.

Nous avons évalué l'impact de l'utilisation d'OFrLex par un étiqueteur en parties du discours. Une telle évaluation est naturellement très limitée, ne serait-ce que parce qu'elle ne fait pas usage du niveau syntaxique du lexique. Nous avons utilisé à cette fin alVWTagger [13], que nous avons développé dans le cadre de notre participation à la campagne d'évaluation CoNLL 2017 dédiée à l'analyse syntaxique multilingue (Villemonte de La Clergerie *et al.*, 2017). Il s'agit d'un étiqueteur statistique qui, comme son prédécesseur MElt (Denis & Sagot, 2012), peut s'appuyer sur un lexique externe pour produire des traits qui viennent en complément des traits extraits du corpus d'apprentissage ou de test. Nous avons utilisé comme données pour l'apprentissage la section d'entraînement de la version *Universal Dependencies* (v2.4, Nivre & *al.*, 2019) du SRCMF, et la section de développement de ce même corpus comme données d'évaluation. Nous avons entraîné avec alVWTagger deux modèles d'étiquetage en catégories UD, l'un sans utilisation d'OFrLex et l'autre en utilisant un lexique de formes fléchies extrait d'OFrLex dans lequel chaque forme est associée à sa partie du discours OFrLex. Les résultats sont probants : l'utilisation d'OFrLex fait passer l'exactitude globale de 93,8% à 94,8%, et l'exactitude sur les seuls mots inconnus du corpus d'entraînement, qui représentent 8,5% des 16 463 mots du corpus d'évaluation, de 81,6% à 85,7%.

5 Discussion

Dans sa version actuelle (version 1), OFrLex n'est pas encore une ressource fiable, malgré sa large couverture. Au niveau morphologique, les classes flexionnelles fournies sont fiables, mais certains verbes irréguliers sont encore imparfaitement décrits. Les informations sémantiques (gloses, définitions, liens vers les entrées des ressources de départ) sont assez fiables, mais certains exemples à revoir ont déjà été identifiés. Enfin, au niveau syntaxique, l'approche décrite ci-dessus est rudimentaire : elle n'a pour vocation que de produire une première version du lexique, version dont l'existence permet d'utiliser des techniques d'amélioration par l'utilisation d'OFrLex dans un analyseur syntaxique.

En effet, le développement d'OFrLex se fait dans un contexte plus large, et va de pair avec celui d'un analyseur syntaxique à large couverture qui s'appuie sur les informations lexicales morphologiques et syntaxiques qu'il fournit. Il s'agit d'un analyseur hybride inspiré de l'analyseur FRMG du français contemporain (Villemonte de la Clergerie, 2005) et dont les premières étapes du développement sont décrites par (Regnault, 2019). Comme FRMG, cet analyseur s'appuie sur une métagrammaire développée à la main, compilée automatiquement en une grammaire d'adjonction d'arbres (TAG) factorisée et avec contraintes. À partir de cette grammaire est produit un analyseur syntaxique symbolique ambigü, complété par un mécanisme de désambiguïsation heuristique ou appris automatiquement à partir d'un corpus arboré. L'analyse automatique de textes par cet analyseur syntaxique pour l'ancien français permettra d'identifier des erreurs et des manques dans OFrLex, soit en comparant des analyses produites avec des analyses de référence lorsque c'est possible, soit par exemple au moyen de techniques de fouille d'erreurs telles que proposées par Sagot & Villemonte de La Clergerie (2008).

Remerciements

Ce travail a été réalisé dans le cadre du projet ANR-16-CE38-0010 PROFITEROLE (2017–2020) dirigé par Sophie Prévost.

13. https://gitlab.inria.fr/almanach/alTextProcessing/alAnalyser

Références

BOHNET B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, p. 89–97, Beijing, Chine.

DENIS P. & SAGOT B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, **46**(4), 721–736.

GUIBON G., TELLIER I., CONSTANT M., PRÉVOST S. & GERDES K. (2014). Parsing Poorly Standardized Language Dependency on Old French. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, p. 51–61, Tübingen, Allemagne.

GUIBON G., TELLIER I., PRÉVOST S., CONSTANT M. & GERDES K. (2015). Analyse syntaxique de l'ancien français : quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage ? In *Actes de la 22ème conférence sur le Traitement Automatique du Langage Naturel (TALN)*, Caen, France.

GUILLOT C., HEIDEN S. & LAVRENTIEV A. (2017). Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques. Revue de Linguistique française diachronique*, **7**, 168–184.

MARTINEAU F. (2008). Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus*, **7**.

NIVRE J. & *al.* (2019). Universal dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

REGNAULT M. (2019). Adapting a Metagrammar for Contemporary French to Medieval French. In *TALN-RECITAL 2019 - 26e édition de la conférence TALN (Traitement Automatique des Langues Naturelles) et 21e édition de la conférence jeunes chercheur×euse×s RECITAL*, Toulouse, France.

SAGOT B. (2010). The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

SAGOT B. (2014). DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German. In *Language Resources and Evaluation Conference*, Reykjavik, Islande : European Language Resources Association.

SAGOT B. & VILLEMONTE DE LA CLERGERIE É. (2008). Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. *Traitement Automatique des Langues*, **49**(1), 41–60.

SAGOT B. & WALTHER G. (2013). Implementing a formal model of inflectional morphology. In C. MAHLOW & M. PIOTROWSKI, Eds., *Third International Workshop on Systems and Frameworks for Computational Morphology*, volume 380 of *Communications in Computer and Information Science*, p. 115–134, Berlin, Allemagne : Humboldt-Universität Springer.

STEIN A. (2014). Parsing heterogeneous corpora with a rich dependency grammar. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Islande.

A. STEIN & AL., Eds. (2008). *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*. Stuttgart, Allemagne : Institut für Linguistik/Romanistik.

STEIN A. & PRÉVOST S. (2013). Syntactic annotation of medieval texts : the Syntactic Reference Corpus of Medieval French (SRCMF). In P. BENNETT, M. DURRELL, S. SCHEIBLE & R. WHITT, Eds., *New Methods in Historical Corpus Linguistics*, Corpus Linguistics and International Perspectives on Language, p. 275–282. Narr Verlag.

VILLEMONTE DE LA CLERGERIE E. (2005). DyALog : a tabular logic programming based environment for NLP. In *Proceedings of 2nd International Workshop on Constraint Solving and Language Processing (CSLP'05)*, Barcelone, Espagne.

VILLEMONTE DE LA CLERGERIE É., SAGOT B. & SEDDAH D. (2017). The ParisNLP entry at the ConLL UD Shared Task 2017 : A Tale of a #ParsingTragedy. In *Conference on Computational Natural Language Learning*, Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies, p. 243–252, Vancouver, Canada.