OFrLex: A Computational Morphological and Syntactic Lexicon for Old French

Gaël Guibon, Benoît Sagot

OFrLex: A Computational Morphological and Syntactic Lexicon for Old French

Gaël Guibon, Benoît Sagot
LLF (CNRS) – Université Paris Diderot & Almanach (Inria),
Almanach (Inria)
Paris, France
{firstname.lastname}@inria.fr

Abstract

In this paper we describe our work on the development and enrichment of OFrLex, a freely available, large-coverage morphological and syntactic Old French lexicon. We rely on several heterogeneous language resources to extract structured and exploitable information. The extraction follows a semi-automatic procedure with substantial manual steps to respond to difficulties encountered while aligning lexical entries from distinct language resources. OFrLex aims at improving natural language processing tasks on Old French such as part-of-speech tagging and dependency parsing. We provide quantitative information on OFrLex and discuss its reliability. We also describe and evaluate a semi-automatic, word-embedding-based lexical enrichment process aimed at increasing the accuracy of the resource. Results of this extension technique will be manually validated in the near future, a step that will take advantage of OFrLex's viewing, searching and editing interface, which is already accessible online.

Keywords: Morphological lexicon, Syntactic lexicon, Lexicon Enrichment, Old French

1. Introduction

Old French regroups romance languages qualified as Oïl languages used in the north of France, south of Belgium and in the Anglo-Norman islands spoken from 8th century to 14th century. They contrast with the Oc languages that come from the south of France. Contrary to Middle French, Old French possesses nominal declination. Both led to contemporary French and possess relatively free word order: verbs are often in second position following a non subject constituent. Moreover, there is no spelling standardisation in Old French, even for proper nouns from the same author. The main textual databases with semi-automatic lemmas and part-of-speech tags (PoS) are the *Base de Français Médiéval* (BFM - Medieval French Base) (Guillot et al., 2017)[1] with more than 4 million words and the *Nouveau Corpus d'Amsterdam* (NCA - New Amsterdam Corpus) (Stein and al., 2008)[2] with more than 3 million words. The main treebanks for Old French are the Syntactic Reference Corpus of Medieval French (SRCMF) (Stein and Prévost, 2013) and the Old French subpart from the *Modéliser le changement : les voies du français* (MCVF) corpus (Martineau, 2008). However, they do not share the same syntactic and POS tag sets, and only SRCMF is on open access with part of it in Universal Dependencies (UD) (McDonald et al., 2013) format[3].

In the available resources different kinds of text are gathered. Some vary in style (prose, verse), literary genre (religious, historical, didactical, *etc.*), or even in time span (from 10th century to 13th century). Nevertheless, there is no available morphological lexicon,[4] and *a fortiori* no syntactic lexicon[5] for Old French. Most of the existing lexicons and dictionaries are either not made for later natural language processing exploitation or only contains minimal morphological (and sometimes syntactic) information.

In this paper, we present the morphological and syntactic lexicon for Old French named OFrLex. The creation of this lexicon is semi-automatic with a substantial manual process. Moreover, it forced the resolution of multiple obstacles: to structure and merge multiple resources not necessarily originally structured, to fuse the heterogeneous and not always consistent lexical information, and to create lexicon information such as morphological classes and valency from scratch or from incomplete source information. Hence, OFrLex was made using automatic tools and manual correction or addition of information. This lexicon can be used for improving Old French dependency parsing and PoS tagging.

The paper is organised as follows. We start by summarising related work (Section 2.) before presenting the lexicon initial creation process with the different language resources used (Section 3.3.). We then present our methodology to automatically enrich the lexicon (Section 4.2.) and explain our distribution strategy for OFrLex (Section 5.). Finally, we show preliminary results on PoS tagging using the lexicon (Section 6.) before tackling future work and improvements (Section 7.).

[1] http://bfm.ens-lyon.fr

[2] https://sites.google.com/site/achimstein/research/resources/nca

[3] https://github.com/UniversalDependencies/UD_Old_French-SRCMF/

[4] A morphological lexicon is a collection of entries of the form ⟨inflected form, lemma (often a citation form), morphological features⟩ (extensional inflectional lexicon) or a collection of entries of the form ⟨citation form, inflection class label⟩ associated with an inflectional grammar that defined how to generate inflected forms given the citation form and an inflection class label (intensional inflectional lexicon).

[5] A syntactic lexicon associates each entry (generally at the lexeme level) with syntactic information, including valency information, control/raising/attribution information, and other types of information describing the syntactic behaviour of the entry.

2. Related Work

Recent work used the previously mentioned textual databases for Natural Language Processing (NLP) tasks. PoS tagging has been applied on SRCMF using TreeTagger (Schmid, 1999; Stein, 2014) and Conditional Random Fields (Lafferty et al., 2001; Guibon et al., 2014; Guibon et al., 2015) as a preparation for Old French dependency parsing using Mate (Bohnet, 2010).

On the other hand, lexicon enrichment is a part of the lexicon creation process and has been the subject of several research work, particularly for morphological lexicons. Nicolas et al. (2010) developed an unsupervised morphological rule acquisition tool which was combined with the Alexina framework (Walther and Nicolas, 2011; Nicolas et al., 2012) to enrich morphological lexicons. Another approach used to enrich or create a lexicon is derived from *parsebanking* (Rosén and de Smedt, 2007) which consists of creating a new treebank by applying a well-known and tested grammar or parser on the corpus. Recently, incremental *parsebanking* showed good results for enriching morphological lexicons with high coverage (Rosén et al., 2016). Valency retrieval through deverbative nouns was also tackled (Fučíková et al., 2016) but requires a task oriented gold dataset. Another recent enrichment strategy consists into using word embeddings to obtain clusters of words in order to enrich a lexicon (Siklósi, 2016).

Morphological lexicons have been used for several tasks. From constraints derived from lexicon at PoS tagging time (Kim et al., 1999; Hajič, 2000) to additional lexicon-based features combined with standard ones during the training process (Chrupała et al., 2008; Goldberg et al., 2009; Denis and Sagot, 2012). To improve these lexicon usages for different tasks such as multilingual PoS tagging supported by a lexicon (Sagot, 2016), we need to create a computational morphological lexicon for Old French: the OFrLex lexicon.

3. Lexicon Creation

3.1. Heterogeneous Language Resources

The idea behind OFrLex is to derive all information from different sources in order to obtain a morphological Old French lexicon. We try to take into consideration all freely available language resources for this task.

FROLEX With this objective in mind we first used FROLEX[6] (Serge Heiden, 2016). The FROLEX lexicon is a combination of information coming from the *Base de Français Medieval* (BFM - Medieval French Base) (Guillot et al., 2017), the *Nouveau Corpus d'Amsterdam* (NCA - New Amsterdam Corpus) (Stein and al., 2008), and the *Dictionnaire du Moyen Français* (ATILF, 2015) (DMF - Middle French Dictionary). These language resources being already merged in one resource, we use the million extensional entries from FROLEX. By extensional entry, we refer to the fact that each one of these entries links to an attested inflected form, and not a lexeme, as visible in Table 1. Depending on the sources, information for each entry may vary. However, the part-of-speech tags (PoS) are already converted to their CATTEX[7] (Guillot et al., 2010) equivalent with additional gender and number. Even if this resource is convenient as it merge multiple ones, some of the entries have noise (i.e. multiple entries for one form with same incomplete information). Moreover, lemmas do not follow the same convention depending on the source from which they were extracted. The usage of DMF, a dictionary for Middle French, and the fact that lemmas are not represented by all their inflected forms, makes some entries and silence irrelevant for our purpose of obtaining a morphological lexicon for Old French.

Wiktionary Wiktionary[8] is a free dictionary which contains 6,500 entries for Old French corresponding to a lexeme and containing formalised descriptions for the inflection classes. The lexeme *mengier*[9] (i.e. to eat) comes with alternative forms such as *mangier*, along with the etymology, and english gloss, and inflection information. We use the extraction process described in Sagot (2014): converting Wiktionary (wiki format) into a structured XML file before using it to extract morphological entries. A morphological entry consists of a citation form, an inflection class identifier, and the list of stems or irregular forms if relevant. Finally, we manually developed a morphological grammar describing the most important inflection classes present in Wiktionary. This morphological grammar use the Alexina format (Sagot and Walther, 2013). For instance, for verbs we use a model containing 8 stems and 2 exponent levels: an intermediate level for some consonant palatalisation at the end a stem for instance, and higher level for standard terminations in 4 set of rules. The latter follows the Paradigm Function Morphology principle (Stump, 2006).

***Altfranzösisches Wörterbuch* by Tobler and Lommatzsch (TL)** *Altfranzösisches Wörterbuch* (shorten as TL) is the reference dictionary for Old French, written in German. We used two versions created and distributed by Peter Blumenthal and Achim Stein[10].

- The first version is made of a list of lemmas manually obtained accompanied by an index of forms from the Godefroy's dictionary. Each information possesses a source information "tl" for TL and "g" for the Godefroy's dictionary. Simplified entries are visible in Table 2. In this Table, main entries (*Haupteingtrag*) are distinguished from secondary entries or variants (mainly graphical ones). Secondary entries are linked to the main one in a many-to-one fashion. Moreover, multiple reference links are given for main and secondary entries: page, line, *etc.*
- The second version used is obtained through Optical Character Recognition (OCR) with numerous recog-

[6] https://github.com/sheiden/Medieval-French-Language-Toolkit

[7] CATTEX is a set of Part-of-Speech tags taking into account morphosyntactic information.

[8] https://en.wiktionary.org/

[9] https://en.wiktionary.org/wiki/mengier#Old_French

[10] https://www.ling.uni-stuttgart.de/institut/ilr/toblerlommatzsch/downloads.htm

Form	Frequency		Original tag			Extended CATTEX tag		Lemma	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: Example entries from FROLEX

nition errors. We partially corrected this version manually by focusing on the important parts such as the type of the word. Table 3 presents an example of this manual correction. We then automatically extracted informations by first checking form errors and ignoring the entry if we found any. An example of the extraction result is visible in the bottom part of Table 3.

***Lexique de l'ancien français* by Godefroy** We consider the Wikisource version of the *Lexique de l'ancien français* (Old French Lexicon)[11]. Figure 1 shows the online version used. This resource has already been made by applying OCR over the original text and then partially correcting it. It possesses a wide coverage albeit with ghost words and meanings. These ghost words are lexical units wrongly considered as such. Thus, we filtered it using the dedicated ghost words base named *Base des mots fantômes [du Godefroy]*[12] dedicated to identify these entries and to clean them. Moreover, this lexicon covers up to the XV century, which is not Old French anymore but Middle French. This data being structured, we easily extracted citation forms, CATTEX PoS tags with additional gender if relevant, a definition, and the link to the corresponding page.

- **aaisant**, adj., commode.
- **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: Godefroy's lexicon from Wikisource

***Dictionnaire Électronique de Chrétien de Troyes* (DECT).** The dictionary by Chrétien de Troyes was written during the 12th century and is distributed by the CNRTL[13] in a PDF format (DECT). We converted it in a textual format and extracted entries in a semi-automatic fashion using simple rules. This resource is useful because it links entries with other dictionaries such as TL and Godefroy. Inflected forms are also available for each entry.

[11] https://fr.wikisource.org/wiki/Lexique_de_l'ancien_francais

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

[13] French national center of textual and lexical resources: https://www.cnrtl.fr/

3.2. Merging information

To create the OFrLex lexicon we need to aggregate all sources by linking information to unique entries. To do so, we first use the citation forms contained in TL using all DECT entries and their explicit reference to TL entries. Very few errors were found during this process. However, to obtain a large coverage we also use other sources when lemmas linked to multiple matches from Godefroy, TL, and/or DECT. If a lemma differ from one source to another, we create multiple entries and disambiguate them manually based on the definitions obtained from other resources. However if the lemma is the same we fuse their information.

Morphology. Morphological features such as gender, number, person, tense and mood are extracted from Wiktionary entries in a semi-automatic way. Indeed, if the citation form is available we retrieve information automatically from sources. If it is not available we add it manually when possible.
Form variants are associated based on FROLEX entries.

Result. By applying this semi-automatic process, we obtain a morphological lexicon where one entry (i.e. one lexeme) is linked to the different sources. This lexicon also contains information derived from glosses, definitions and variants from the different sources. Note that the Universal Part-of-Speech (UPoS, i.e. the UD morphological category) is also extracted from these sources by converting each different PoS tags into a matching UPoS. Table 4 show quantity information from OFrLex per UPoS.

3.3. Syntactic Information Addition.

We complete this morphological lexicon for Old French with syntactic information. To do so, we follow the Alexina conventions already used for the contemporary French morphological lexicon L*efff* (Sagot, 2010). From L*efff* we obtain different types of syntactic information such as redistribution and valency. To retrieve them we make the hypothesis that verbs syntactically similar between Old French and Contemporary French can share information if and only if the former do not possesses any in the lexicon. To be more precise, we use different types of information from L*efff* with a hierarchical priority presented in Algorithm 1.
In this process, valency is retrieved from multiple sources (Godefroy, TL and DECT) looking for textual markers such as "I" (intransitiv in TL), "trans." (transitiv) or "refl." with multiple spelling variants.
Finally, Table 5 presents 3 entries from OFrLex: "afiner", "afiner$_2$" and "effiner". Those entries are Old French vari-

Lemma *Lemma*	Haupt-eintrag *Main entry*	Wortart *Category*	Var. *Variant*	Werk *Source*	Band *Volume*	Spalte *Page*	Zeile *Line*	IstVar. *Is a variant*
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: Examples from Tobler-Lommatzsch entries index

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 (. . .)	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 (. . .)

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: Example from Tobler-Lommatzsch (OCR version) before (left panel) and after (right panel) partial correction. Bottom panel shows extracted structured entries

UPoS	**#lexemes (intentional entries)**	**#extentional entries**
ADJ	7,878	75,490
ADV	1,843	4,182
CCONJ	37	222
SCONJ	52	303
DET	156	1,963
INTJ	203	1,003
NOUN	44,063	133,664
PROPN	1,944	10,501
ADP	286	1,965
PRON	476	2,685
VERB	16,784	583,854
PUNCT	19	41

Table 4: OFrLex intentional and extensional entries for each UPoS.

ants for the contemporary French verb "affiner" (i.e. to refine). This example show two entries for the same spelling differentiated by their syntactic information. The third row is the variant of the second row "afiner$_2$" and serves to decipher the encapsulated syntactic information from the Alexina standard to natural language. In Alexina, information are separated in two parts. First the citation form (in bold), the inflection class (*v-er*) and the syntactic information are visible as tabulated separated values (TSV). Then, unlimited comments can be added after the hashtag (#) for meta-information such as a link, source of information or variants. These information are XML encoded, using tags as categories and attributes as detailed information (Sagot, 2010).

Algorithm 1 Semi automatic syntactic information retrieval

```
for all OFrLex entries do
  if Lefff & OFrLex similar syntactic infos then
    Add "Pseudo gloss"
  else if Contemporary french gloss available then
    Add French Gloss (from sources or manually)
  else if French descendent available then
    Add French Descendent (Wiktionary or manually)
  else if Lefff citation form == OFrLex citation form then
    Dispatch syntactic informations
  else if Valency available then
    Valency from Godefroy or TL or DECT
  else
    Set default value to "transitive"
  end if
end for
```

4. Lexicon Enrichment

Once created, the lexicon is not reliable enough to be used as a reference source of information for Old French. We need to enrich it and validate it by Old French or diachrony specialists. However, the manual process is long and tedious especially when it comes to enrich the lexicon by decreasing the silence rate. This is why we first automatically enrich the lexicon before planning on the validation phase. The objective is now to obtain additional information for the lexicon. We derive from Siklósi (2016) by being stricter: the user interface (UI) only allows to dispatch information. Plus, we focus on variant candidates for a non spoken language. This is why this additional information is not necessarily expected to be correct. We make the following hypothesis: it is easier and faster to correct or validate some information and errors than trying to find missing values from scratch, especially when dealing with a non living language with relatively few experts and resources.

4.1. Information from variants

Valency Valency information was initially obtained from Lefff for verbs that were considered syntactically similar, or by manual insertion (see Section 3.2.). Nevertheless, not all verbs were covered and other categories were not taken into account in the process. This is why we computed valency information from the variants found for each entry. *In fine* valency can come from the manual valency inserted, the gloss, the pseudo-gloss, Godefroy or Tobler definitions, or from itself (e.g. if it is directly available in DECT).

Lemmas When a lemma is missing, we compute candidate lemmas by analysing variants in a two-way fashion by

afiner$_1$ v-er 100;Lemma;v;<Suj:cln|sn,Obj:(cla|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, first group verb with regular inflection Passive transitive verb with nominal subect or clitic. And optional direct object, nominal or clitic Variant of "afiner"$_2$ Corresponding entry from Tobler-Lommatzsch: afiner2 (1:189:47) '[trans.] läutern' Valency infered from the Tobler-Lommatzsch gloss

Table 5: OFrLex syntactic information: entries with one more explicit version ("effiner")

.

populating and spreading the same information across variants and the original lexeme. To deal with multiple lemmas we decided to take the first lemma found for the same category. If there is none we take another one randomly.
This information computed from variants and distant variants—the variant of a variant or the variant of a gloss—will not be used as reference but as pre-annotations provided to human validators.

4.2. Generation of pseudo-synonyms

Lemmas and valency information were obtained using variants, making the process of dispatching information between entries a relevant strategy. In order to continue using this approach, we aim to generate variant candidates that we name pseudo-synonyms. These pseudo-synonyms are not necessarily morphological variants but can find their similarity in morphology, spelling or sense. Our objective is to propose possible enrichment automatically obtained that the user will be able to validate or refute.
To automatically obtain pseudo-synonyms we need to consider the words in context given their morphosyntactic category. This is why we use the BFM corpus in its two available versions: 170 raw texts and 42 CONLL files with verified part-of-speech (PoS) tags. The new annotated BFM corpus which is the current biggest annotated corpus for Old French. Table 6 shows information about the PoS tagged version.

Tokens	3,640,013
Vocab	158,620
EN POS tags	20
FR POS tags (CATTEX)	65

Table 6: Data used for candidates

We start by using the raw texts as input to train a FastText model (Joulin et al., 2017) using the Gensim implementation[14]. FastText was selected for various reasons. First, we need to take into account morphological information about the words with their inflections. Formal similarity could be used externally but the bag of n-grams used in this model is already dedicated to this. Second, we do possess relatively small data (see Table 6) in comparison to other languages with a lot of resources. Thus, we cannot use latest models such as Bert (Devlin et al., 2019) or ELMo (Peters et al., 2018) which require a large amount of data. In fact, we tried both architectures of Word2Vec (Mikolov et al., 2013) with inconclusive results.
In our methodology we need to distinguish inflectional forms (f) with lexemes (L), the former are obtained from raw text while the latter are extracted from OFrLex lexicon. To obtain a lexeme embedding ($\mathbf{e}(L)$), we apply the FastText model trained on the raw text corpus made of inflectional forms ($\mathbf{ft}(f)$). We then make the average of the form embeddings from the lexeme, weighted by the occurrences of each form with the same the PoS tag (p) as the lexeme. The weighted average has recently been demonstrated to be a good approach to obtain meta-embeddings (Coates and Bollegala, 2018). Here we apply this logic while taking into account occurrences per PoS tag. This is formalised in Equation 1.

$$\mathbf{e}(L) = \frac{\sum_{f \in F(L)} \mathbf{ft}(f)\,\mathrm{occ}(f)}{\sum_{f \in F(L)} \mathrm{occ}(f,p)} \quad (1)$$

We use the set of lexeme embeddings obtained using equation 1 as an input for clustering. We cluster this lexeme embedding space using Spectral Clustering (Ng et al., 2002). As for the hyper-parameters we set a Gaussian kernel, a gamma of 0.7 and discretisation to assign clusters. Moreover, we do not use eigenvalue decomposition strategy and set the number of targeted clusters as 20, according to the number of PoS tags (n clusters = n distinct PoS tags). These predicted clusters are meant to be used as an additional verification for pseudo-synonyms, but cannot be evaluated as we do not possess gold labels for them.
Once we have the lexeme embeddings and their predicted cluster, we can obtain the nearest neighbours for each lexeme. Given a lexeme, we take all the other lexemes with the same PoS tag and that share the same cluster. We then compute the cosine distance between their embeddings (equation 2). This process is formalised in equation 3 where

[14] https://radimrehurek.com/gensim/

$C(L_i)$ is the cluster for the given lexeme.

$$K(\mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{||\mathbf{e}_i||.||\mathbf{e}_j]]} \quad (2)$$

$$\forall L_i, \forall L_j \in C(L_i), d(L_i, L_j) = 1 - K\left(\mathbf{e}(L_i), \mathbf{e}(L_j)\right) \quad (3)$$

Finally, nearest neighbours (nn) are obtained by keeping the lexeme with the minimum cosine distance with the targeted lexeme ($d(L_i, L_j)$), visible in Equation 4. The resulted nearest neighbour is controlled by the clustering, the formal similarity induced by the bag of n-grams, and the lexemes' PoS tag. Because it is not exactly a nearest neighbour nor a variant, we name it pseudo-synonym.

$$\mathrm{nn}(L_i) = \underset{L_j \in C(L_i), j \neq i}{\arg\min} \; d(L_i, L_j) \quad (4)$$

By applying this methodology we get a pseudo-synonym for 15,041 lexemes out of 54,087. Thus, only 27.81% of lexemes obtain a pseudo-synonym, i.e. a candidate for possible source information retrieval. These pseudo-synonyms are then used as propositions for the validator.

It is not an easy task to automatically evaluate these pseudo-synonyms but we try to give a glimpse of its quality and usage by taking 10 random pseudo-synonyms found and verifying manually their soundness. For each pseudo-synonym we want to know if it is a probable variant or if it already exists as a variant in OFrLex. Table 7 shows this information in the first two columns, followed by the UPoS, the source lexeme from which we want to find pseudo-synonyms, and the pseudo-synonym (*v1*) followed by nested pseudo-synonym – pseudo-synonym of the preceding pseudo-synonym –.

The first pseudo-synonyms are *voutroillier* and *voutroiier* for the lexeme *voutrillier* (i.e. se vautrer – to wallow –). This pseudo-synonym is a correct candidate for information extraction as this variant occurrence is already known from OFrLex and was extracted from TL. The second, *menestralsie* for the lexeme *menestraucie* (i.e. act of production of a minstrel) is a new graphical variant which we can verify by manually looking at the definition of *menestraudie* from DMF which reports them. Plus, they follow the same genre and valency. For the lexeme *auberc haubert* (coat of chain mail) multiple spelling variants can be found in text as reported in the Littré[15], this pseudo-synonym is a correct candidate for information dispatchment.

We apply the same manual checking for each one and showed that 60% are correct in this small non representative subset. However, among the 4 ones that we do not seem to find any proof for, two pseudo-synonyms are probable good candidates considering their form (*gaagnier*) or the quite similar definitions (*mas* is related to arable ground where *massiz* and *massëiz* define something made of the same material).

In any case, we use the pseudo-synonyms as a support (*proposition*) for validators to find or discover new variants and finally enrich the lexicon, and not to directly insert it in OFrLex without validation.

[15] https://www.littre.org/definition/haubert

5. Distribution and Improvements

Language resources used to create OFrLex are either free (DMF is free for non commercial usage), from public domain (Godefroy's lexicon and dictionary), or follow a copy left pattern - BFM, SRCMF, FROLEX and Wiktionary follow a CC BY-NC-SA[16]). Hence, we follow the same licence and distribute OFrLex from its git repository: https://gitlab.inria.fr/almanach/alexina/ofrlex.

The OFrLex repository possesses files for the intentional lexicon (Alexina~~FATRSLI~~ format) and the extensional lexicon. The latter is the ready-to-use lexicon with all entries and their inflected forms automatically derived from the set of inflectional rules contained in the intensional lexicon. Table 8 shows an example for the "afiner" entry in the intensional lexicon and 2 of the many inflected forms derived from it for the extensional lexicon.

OFrLex is created semi-automatically and requires thorough validation by Old French specialists. To deal with this issue and to facilitate validation, we developed a user interface dedicated to OFrLex edition and validation. All modifications made in the interface will be automatically integrated in future versions of OFrLex following the architecture shown in Figure 2. Fully automatic lexicon enrichment such as pseudo-synonyms (see Section 4.2.) and valency or variants information are indicated in the interface as "propositions." They are not fully integrated in the OFrLex source database but are visible for the annotator which can validate them, thereby triggering their integration. This relies on the distinction between 3 information types distinguished by colours: validated information, semi-automatic information, and propositions (with source/confidence indicator). This web interface also serves as a search engine (at the lexeme level) via the public API.

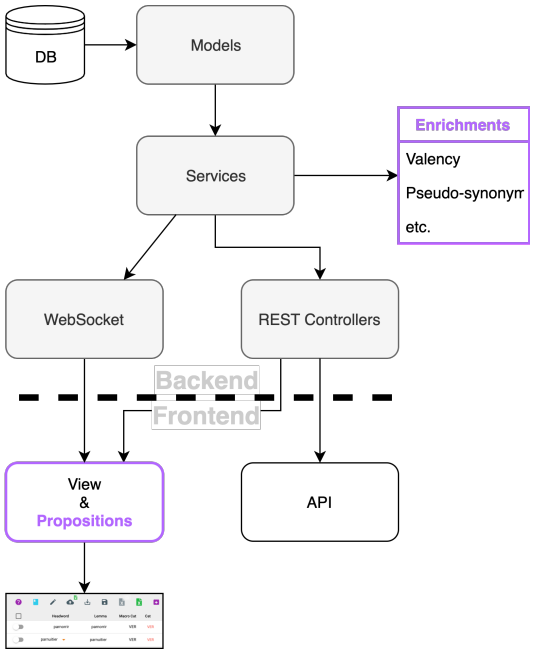

Figure 2: User Interface architecture for OFrLex validation

[16] https://creativecommons.org/licenses/by-nc-sa/2.5/

Variant	Existing	POS	Lexeme	v1	v2
True	True	VERB	voutrillier	voutroillier	voutroiier
True	False	NOUN	menestraucie	menestralsie	
True	False	NOUN	auberc haubert	auberc aubert	
False	False	VERB	articulariser	articuler	
True	False	NOUN	emblavëure	emblaëure	
False	False	NOUN	mediqué	mediomatrique	
True	False	VERB	foibloiier	forploiier	
False	False	VERB	gaaingnier	gaagnier	
True	False	VERB	entrenväir	entrenvair	
False	False	ADJ	mas	masi	massëiz

Table 7: Subset of random pseudo-synonyms for manual inspection.

Intensional Lexicon (Lexemes with inflectional and syntactic information)

afiner$_1$ v-er 100;Lemma;v;<Suj:cln|sn,Obj:(cla|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"/>

Extensional Lexicon (inflected forms generated from intensional entries)

afinent$_1$ v 100
pred="afiner___60674__1<Suj:cln|sn,Obj:(cla|sn)>",@pers,cat=v,upos=VERB,@pl.3.subj.prs.std
afinera$_1$ v 100
pred="afiner___60674__1<Suj:cln|sn,Obj:(cla|sn)>",@pers,cat=v,upos=VERB,@sg.3.ind.fut.std

Table 8: OFrLex intensional and extensional examples

.

6. Preliminary Usage

This lexicon can be used for multiple purposes such as diachronic studies, dependency parsing for Old French or PoS tagging. We evaluated OFrLex impact on PoS tagging using the Universal Dependencies (Nivre and *al.*, 2019) version of SRCMF treebank's training set. In order to do so we trained three models using alVWTagger[17] initially developed for the CONLL-2017 shared task (Villemonte de La Clergerie et al., 2017). Like MElt (Denis and Sagot, 2012), this PoS tagger can use an external lexicon to infer complementary information from the train set or the test set. Thus, we only use OFrLex to extract the inflected forms with their associated PoS tag as the external lexicon for one model, and no external resource for the second model.

Model	Accuracy	Unknown words Accuracy
alVWTagger	93.80	81.60
alVWTagger + OFrLex v1	94.80	85.70
alVWTagger + OFrLex v1.2	**95.08**	**87.10**

Table 9: PoS tagging accuracy scores on SRCMF-UD using alVWTagger combined with the initial OFrLex (v1) and the one currently under enrichment and validation (v1.2).

The first model using OFrLex improved the global accuracy from 93.8% to 94.8%. More importantly, unknown words accuracy increased by 4 points: from 81.6% to 85.7%. This improvement on the 16,463 unknown words (8.5% of the test set) supports the need of a dedicated lexicon for NLP tasks. Moreover, we also trained alVWTagger with OFrLex after validation and enrichment using the interface. This led to an improvement in accuracy, both overall and on unknown words. This promising results motivates the need for an incremental validation phase helped by automatic suggestions. Of course, this represents only a small task and cannot be enough to fully take advantage of OFrLex which contains more information than just the PoS tags. However it serve as a preliminary example of its use.

[17] https://gitlab.inria.fr/almanach/alTextProcessing/alAnalyser

7. Future Work

In this paper we presented the OFrLex creation process to obtain a morphological and syntactic lexicon for Old French from heterogeneous resources, along with the methodology used to enrich it, taking into account the fact that it is not a living language. For the moment, syntactic information is mostly limited to verbs; why we plan on extending it to adjectives and nouns in the near future. As shown in Section 5., the user interface is currently used for the lexicon validation phase supported by multiple enrichment propositions, such as those described in Section 4.2..

Even if our preliminary results focused on part-of-speech tagging, we plan to also use *parsebanking* as a way to improve the lexicon. To do so, a meta grammar for Old French parsing is under development (Regnault et al., 2019) and already uses OFrLex to improve parsing quality and to incrementally fix possible noise or silence present in the lexicon.

OFrLex is available for everyone and future validation will yield new versions once the validation phase is done.

Acknowledgements

This work was partly funded by the French national ANR grant PROFITEROLE (ANR-16-CE38-0010) headed by Sophie Prévost, as well as by the second author's chair in the PRAIRIE institute,[18] funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

Bibliographical References

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, pages 89–97, Beijing, Chine.

Chrupała, G., Dinu, G., and Van Genabith, J. (2008). Learning morphology with morfette. *Sixth International Conference on Language Resources and Evaluation (LREC)*.

Coates, J. and Bollegala, D. (2018). Frustratingly easy meta-embedding – computing meta-embeddings by averaging source word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 194–198, New Orleans, Louisiana, June. Association for Computational Linguistics.

Denis, P. and 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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Fučíková, E., Hajič, J., and Urešová, Z. (2016). Enriching a valency lexicon by deverbative nouns. In *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)*, pages 71–80, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Goldberg, Y., Tsarfaty, R., Adler, M., and Elhadad, M. (2009). Enhancing unlexicalized parsing performance using a wide coverage lexicon, fuzzy tag-set mapping, and em-hmm-based lexical probabilities. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 327–335. Association for Computational Linguistics.

Guibon, G., Tellier, I., Constant, M., Prévost, S., and Gerdes, K. (2014). Parsing Poorly Standardized Language Dependency on Old French. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 51–61, Tübingen, Germany.

Guibon, G., Tellier, I., Prévost, S., Constant, M., and Gerdes, K. (2015). Searching for discriminative metadata of heterogenous corpora. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 72, Warsaw, Poland.

[18] `http://prairie-institute.fr/`

Guillot, C., Prévost, S., and Lavrentiev, A. (2010). Manuel de référence du jeu cattex09. *technical manual, UMR ICAR, CNRS/ENS-LSH< http://bfm. ens-lyon. fr/article. php3*.

Guillot, C., Heiden, S., and 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.

Hajič, J. (2000). Morphological tagging: Data vs. dictionaries. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 94–101. Association for Computational Linguistics.

Joulin, A., Grave, É., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Kim, J.-D., Lee, S.-Z., and Rim, H. C. (1999). Hmm specialization with selective lexicalization. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856.

Nicolas, L., Farré, J., and Molinero, M. A. (2010). Unsupervised learning of concatenative morphology based on frequency-related form occurrence. In *Proceedings of the Morpho Challenge 2010 Workshop*, pages 39–43.

Nicolas, L., Farré, J., and Darme, C. (2012). Unsupervised acquisition of concatenative morphology. In *The eighth international conference on Language Resources and Evaluation (LREC)*, pages 865–872, Istanbul, Turkey.

Nivre, J. and *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.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Associ-*

ation for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Regnault, M., Prévost, S., and Villemonte de la Clergerie, E. (2019). Challenges of language change and variation: towards an extended treebank of medieval French. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 144–150, Paris, France, 28–29 August. Association for Computational Linguistics.

Rosén, V. and de Smedt, K. (2007). Theoretically motivated treebank coverage. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 152–159, Tartu, Estonia, May. University of Tartu, Estonia.

Rosén, V., Thunes, M., Haugereid, P., Losnegaard, G. S., Dyvik, H., Meurer, P., Lyse, G. I., and De Smedt, K. (2016). The enrichment of lexical resources through incremental parsebanking. *Language Resources and Evaluation*, 50(2):291–319.

Sagot, B. and Walther, G. (2013). Implementing a formal model of inflectional morphology. In Cerstin Mahlow et al., editors, *Third International Workshop on Systems and Frameworks for Computational Morphology*, volume 380 of *Communications in Computer and Information Science*, pages 115–134, Berlin, Allemagne, September. Humboldt-Universität, Springer.

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, May.

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

Sagot, B. (2016). External lexical information for multilingual part-of-speech tagging. *arXiv preprint arXiv:1606.03676*.

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Siklósi, B. (2016). Using embedding models for lexical categorization in morphologically rich languages. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 115–126. Springer.

Achim Stein et al., editors. (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*. Institut für Linguistik/Romanistik, Stuttgart, Allemagne.

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.

Stump, G. T. (2006). Paradigm function morphology. In Keith Brown, editor, *Encyclopedia of Language and Linguistics (2nd ed.)*, pages 171–173. Elsevier, Oxford, United Kingdom.

Villemonte de La Clergerie, É., Sagot, B., and 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, pages 243–252, Vancouver, Canada, August.

Walther, G. and Nicolas, L. (2011). Enriching morphological lexica through unsupervised derivational rule acquisition. In *WoLeR 2011 at ESSLLI: International Workshop on Lexical Resources*, Ljubljana, Slovenia.

Language Resource References

ATILF. (2015). *DMF : Dictionnaire du Moyen Français*. ATILF - CNRS Université de Lorraine.

Guillot, C., Heiden, S., and 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.

Serge Heiden. (2016). *FROLEX*. github.

Achim Stein et al., editors. (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*. Institut für Linguistik/Romanistik, Stuttgart, Allemagne.

Stein, A. and Prévost, S. (2013). Syntactic annotation of medieval texts: the Syntactic Reference Corpus of Medieval French (SRCMF). In P. Bennett, et al., editors, *New Methods in Historical Corpus Linguistics*, Corpus Linguistics and International Perspectives on Language, pages 275–282. Narr Verlag.