Corpus and Models for Lemmatisation and POS-tagging of Old French

Jean-Baptiste Camps, Thibault Clérice, Frédéric Duval, Lucence Ing, Naomi Kanaoka, Ariane Pinche

Journal of Data Mining and Digital Humanities

Corpus and Models for Lemmatisation and POS-tagging of Old French

Jean-Baptiste Camps[1], **Thibault Clérice**[1], **Frédéric Duval**[1], **Lucence Ing**[1], **Naomi Kanaoka**[1], **Ariane Pinche**[1]

[1]Centre Jean-Mabillon, École nationale des chartes, Université Paris, Sciences & Lettres

Corresponding author: Jean-Baptiste Camps , Jean-Baptiste.Camps@chartes.psl.eu

Abstract

Old French is a typical example of an under-resourced historic languages, that furtherly displays an important amount of linguistic variation. In this paper, we present the current results of a long going project (2015-...) and describe how we broached the difficult question of providing lemmatisation and POS models for Old French with the help of neural taggers and the progressive constitution of dedicated corpora.

Keywords

Lemmatisation; POS tagging; Old French; Historic Languages.

I INTRODUCTION

Today, linguistically annotated corpora[1] are at the centre of crucial research issues, as they constitute long-term and reusable data, ensure the reproductibility of a study, and allow the interrogation and exploitation of large-scale textual sets [Mellet and Purnelle, 2002]. Since the 1960s, the annotation of corpora has been automated by linguists, particularly in the field of NLP (Natural Language Processing), for the quantitative exploitation of linguistic phenomena. The identification of the occurrences of a word is thus facilitated by annotation, which allows a more efficient harvesting. These corpora also constitute essential data for dialectometry or stylometry, in particular for author attribution studies [Mellet, 2002] or for the automatic classification of literary genre [Feldman et al., 2009]. Finally, lemmatisation can be integrated into a semi-automatic processing pipeline starting from the digitisation of a manuscript and leading to the production of a critical edition by comparing the texts of different witnesses [Camps et al., 2019]. The range of possibilities makes this data extremely valuable.

Lemmatisation as an academic task was first developed in the study of flectional ancient languages, such as Latin or Greek, because this computerisation of the task already corresponded to the ancient use of concordance tables or indexes [Rouse and Rouse, 1984]. Roberto Busa and his team working on the lemmatisation of the *Summa Theologica* of Thomas Aquinas starting in the 1950s [Busa, 1980] were pioneers in this area. Ten years later, the LASLA[2] set up a system of lemmatisation of its corpora by collecting all information manually and recording this information in computer files. In the 1990's, the emergence of lemmatisers such as *TreeTagger* [Schmid, 1994] made it possible to automatically annotate corpora. These tools were at the time

[1]The process of lemmatisation associates a canonical form with each term in a text, regardless of its inflection.
[2]Laboratoire d'Analyse Statistique des Langues Anciennes, University of Liège.

based on rules and a dictionary to predict the linguistic annotation of a word (token) in a given language.

In the 1980s, linguistic annotation was extended to Old French with the desire to systematically study the language of a document and its dialectal particularities with the constitution of two manually annotated corpora with a set of 225 numeric tags encoding part of speech and other morphological categorie, the first one of charters and the second one of literary texts, *Amsterdam Corpus*, by Anthonij Dees (Vrije Universiteit Amsterdam) and his collaborators, including Piet van Reenen, in order to produce the two atlas of the linguistic forms of Old French [Dees et al., 1980, 1987]. While the Charter corpus has been lost, in the 2000s, the literary corpus of the second atlas was taken over and fully lemmatised and annotated (addition of POS), with the help of the lemmatiser *TreeTagger* for Old French and using lemmas proposed in the *Tobler-Lommatzsch dictionary* [Tobler and Lommatzsch, 2002] by Pierre Kunstmann and Achim Stein to form the *Nouveau corpus d'Amsterdam* (NCA, Kunstmann and Stein [2007]). But, vernacular languages such as Old French bring other challenges related to spelling variations, making it difficult to use a fixed dictionary for lemmas. Indeed, the same word can have a large number of graphic realisations. For example, the term *cheval* can have more than thirty different forms (see Table 1, Camps et al. [2019]), which makes it pretty much inefficient to use a lemmatisation tool based on a pre-established dictionary.

Form	**Freq.**	**Form**	**Freq.**	**Form**	**Freq.**	**Form**	**Freq.**
cheval	785	chevaux	30	ceuax	10	cheuas	2
cheual	375	chivaus	27	cevax	10	keval	2
chevaus	248	cheuax	23	ceuaus	9	chaval	1
ceval	98	chiual	23	chiuau	9	chavaux	1
chevax	92	cevaus	19	cheuaux	8	cheua	1
chevals	84	chevas	19	kevaus	6	cheualx	1
ceual	66	cheuals	14	chevau	5	cheuau	1
cheuaus	65	cevals	12	cevaux	3	chevalx	1
chival	34	chiuaus	11	chivals	3	chiuals	1

Table 1 – spellings for "cheval" in the *Nouveau corpus d'Amsterdam*

Traditional lemmatisers such as *Treetagger* work with rule sets, a lexicon that only recognises lemmas that already exist in its training corpus, and a decision-tree based algorithm. Due to their fixed lexicon which mainly takes into account only flectional variations and the token environment, the results are not as optimal as those of classical Latin or Modern French due to the unstandardised nature of Old French. However, the most recent developments in the fields of lemmatisation and linguistic annotation of vernacular languages and their variety through the use of recurrent neural networks (RNN) make it possible to set up systems capable of handling the linguistic variations in historical languages[3] thanks to their learning and prediction capacity. Thanks to RNNs, the lemmatiser no longer needs to compare the word in the text to a list of lemmas contained in a dictionary, but is able to predict it sequence by sequence (seq2seq), character by character considering word and sentence levels. This is why we are focusing on the use of these technologies and the implementation of a lemmatisation model for Old French using the lemmatiser and POS tagger *Pie* developed by E. Manjavacas [Manjavacas et al., 2019]),

[3]eg. Middle Dutch [Kestemont et al., 2016], Medieval Latin [Kestemont and De Gussem, 2017], Medieval Occitan [Camps and Couffignal, 2017], Early Irish [Dereza, 2019], Middle High German and a variety of other languages [Manjavacas et al., 2019, Schmid, 2019])

which we have trained from manually annotated corpora that have been aggregated over the course of the projects.

The first experimentss on lemmatisation for Old French were launched in 2015 by Jean-Baptiste Camps and Mike Kestemont, on the texts of the *Chanson d'Otinel* [Camps, 2016], using the newly developed Pandora tagger/lemmatiser that used convolutional and recurrent neural networks [Kestemont et al., 2016]. It soon benefited from insertion in the broader LAKME project (Thierry Poibeau, Daniel Stoekl, J.B. Camps et al.). Initially, the first samples of the *Chanson d'Otinel* have been annotated using the TreeTagger parameters for Old French provided by Achim Stein [Stein, s.d.], then manually corrected, and used to train the first Pandora models; then, phases of data prediction, post-correction and model training were iteratively performed in order to build the corpus. A corrected corpus of 50,000 words annotated with lemmas, morphosyntax and flexion was created for Old French and another equivalent corpus, annotated with lemmas, for Occitan (the latter in collaboration with the CORLIG project of Paris-Sorbonne) [Camps and Couffignal, 2017]. Meanwhile, the lemmatiser Pandora was developed.

Thanks to financial support from PSL (IRIS Scripta) and above all from the DIM STCN (Ile-de-France Region), the project continued after the end of the LAKME project, taking the name OMÉLIE (*Outils et méthodes pour l'édition linguistique enrichie*) from 2018. It then scaled up, carrying out two tasks simultaneously:

1. the development of a lemmatisation post-correction tool, allowing fast correction and batch processing of lemmatiser output;
2. a massive growth of the lemmatised corpus, in order to increase lemmatiser scores.

From a computational point of view, the new annotated corpora allow, by a circular process, the training of more efficient models for lemmatisation and annotation, and thus a subsequent faster growth of the corpus. Both the post-correction application and the lemmatiser are language-independent, and particularly suitable for non-standardised language states. The application has been continuously developed to allow collaborative work and, above all, to be used without advanced computer experience. By 2018, the post-correction application Pyrrha [Clérice and Pilla, 2021] was being developed and used, including for educational purposes. This was a major step forward, since this open-source service significantly improved the lemmatisation speed.

The increasing of the training corpus has made continuous progress thanks to important funding. Its cost is high and can be estimated at between 50 000 and 100 000€, not counting the development of the tools. This means that it is necessary to avoid duplicating data as much as possible and to allow interoperability of the data produced by the various lemmatisation projects in Old French or to promote data sharing or exchange. To this end, a workshop was held at the École des Chartes on 28 October 2018, dedicated to lexical lemma repositories in Old French, in order to standardise practices or at least make them interoperable. It is now necessary to develop equivalence tables between lemmas and to implement them in order to share corpora.

Methods implemented in LAKME and then OMÉLIE have contributed to the setting up of corpora and models for other languages, in particular for Old Occitan, but also for Classical French [Cafiero and Camps, 2019, Camps et al., 2020] (see also infra for the Franco-Italian). After a complete reworking and harmonisation of the corpora, a new version of the Old French models is available since the beginning of 2021 via Pyrrha [Clérice and Pilla, 2021] as well as in the form of an API via Deucalion and Pie-extended [Clérice, 2021][4].

[4]Available at: https://dh.chartes.psl.eu/ and https://tal.chartes.psl.eu/deucalion/.

II CORPORA

Dataset	*Source*	Annotators	Morph	N. tokens	genre
Chrestien	Kunstmann 2009	PK, LI	-	252774	romance
Code	Duval and Pastore, in progress	FC, FD, LI, NK	partial	160007	law
DocLing	Glessgen 2016	NBP, NK	full	68317	charters
Geste	Camps 2016 (et varia)	ACC, JBC, LI, NK	full	195303	epic
Lancelot	Ing (in progress)	LI	-	286095	romance
WauchierSConf	Pinche 2021	AP	full	113694	hagiography
Varia			full	56659	mixed
			TOTAL	1 132 849	

Table 2 – Description of the different corpora produced by the project and used in the experiments.

2.1 Sources and history of the corpus

The annotated texts originated, in a first time, from individual editorial (PhD) projects. The first texts to be tagged were the three manuscripts of the *Chanson d'Otinel* [Camps, 2016], followed by other Old French Epics, now part of the *Geste* database [Camps, 2019]. The *Wauchier* [Pinche, 2021] and *Lancelot* [Ing, 2021] corpora also originate in PhD projects and are constituted of new data. The corpus of juridical texts results from an ongoing research project, concerning a set of vernacular translations of the *Corpus juris civilis*, written between c. 1225-1275. It was also created in the context of the production of complete or partial digital editions of certain translations, in order to characterise and compare translation choices.

Apart from fully new data, preexisting datasets have also been aligned with our reference lists: it is the case of the *Chrestien* corpus, originally produced under the supervision of Kunstmann [2009].

Progressively, new texts have been selected and annotated with the goal of expanding the coverage in terms of chronology, regional scriptae and genres. It resulted in the inclusion of a charters corpus, selected from the DocLing with a sampling by scripta/regional variants [Gleßgen, 2016]. New texts keep being added progressively with this focus, with currently a corpus of lyrical poetry constituted by digitising and annotating preexisting editions [Thibaud IV, 1925, Conon de Béthune, 1925, Brulé and Dyggve Petersen, 1951, Doss-Quinby et al., 2001] and a late allegoric verse text by Guillaume de Digulleville.

2.2 Annotation practice and workflow

The annotation workflow uses a set of tools developed as part of an initiative to establish a fully integrated environment for linguistic annotation and post-correction of historical languages.

Our annotation is characterised by its relative complexity, including the analysis of nominal and verbal flexion: pronouns are distinguished (personal, demonstrative, indefinite, interrogative, relative, adverbial, impersonal, cardinal, ordinal) and times, mode and person are specified for verbs. Once annotated with extent models, the texts and their annotation undergo a post-correction phase inside Pyrrha [Clérice and Pilla, 2021].

Tagging post-correction can be done either linearly (from start to finish), or massively, using concordance tables (accessible from the "Search tokens" link). Certain categories of words are

conducive to massive correction by the relative ease of identification of lemmas (NOMcom, NOMpro, Adverbe, Verbs); on the other hand, POS and morpho-syntaxic flexions always require careful analysis of syntax. For now, inside the application lemma+POS+flexion are not handled as a coordinated unit, and consistency has to be verified by the human.

2.2.1 Lemma

The texts were annotated in lemmas, according to the entries in Tobler-Lommatzsch's dictionary (henceforth, TL), with some adaptations. The choice of using TL instead of, for instance, the lemmas of the *Dictionnaire de Moyen Français* [ATILF, 2015, henceforth DMF] was done bearing in mind the linguistic nature of the corpus to annotate (Old French, not Middle French) and was in line with existing tools at the time [e.g. Stein].

The reviewer verifies and manually corrects pre-annotated forms on the Pyrrha platform (and can also identify and correct at the same time the possible wrong base forms generated by the OCR/HTR of the text). For the disambiguation of lemmas (e.g., *ver1*, masc. noun, 'spring' and *ver2*, masc. noun, 'male pig'), it is still necessary to consult the digitised TL [Tobler and Lommatzsch, 2002] to clearly distinguish the homonyms with their definitions, because the lemmas of the TL are not intuitive and the application does not yet link to definitions. Indeed, there are homonyms even for terms that we do not think about.

For instance, in the sentence

> *El* ***mont*** *n' a home de si grant hardement*

mont isn't 'mountain', but *monde1* (<MUNDUS), 'world'. Such examples are very numerous, due to the phonetic evolution of French that creates many homographs, e.g., *errer*, *errer1* < ERRARE, 'to make a mistake' or *errer2* < ITERARE, 'to go, to move'; *mes*, *mes1* < MA(N)SUM, 'house, garden', *mes2* < MESSIS, 'harvest, reaping', *mes3* < MISSUS, 'messenger', *mes4* < MISSUS, 'delicacies'.

The number of homograph forms is yet increased by morphological and graphetic variation, that can cause alternative spellings to collide between lemmas: e.g. *mes* can be *mes1*, *mes2*, *mes3* or *mes4*, but it can also be a form of *mais1* < MAGIS, 'more', or of *mais2*, 'bad', *mois*, 'month', *manoir*, vb. 'to stay', *metre2*, vb. 'to put, to set', *mon*, poss. 'my',...

In general, we validate or correct the lemmas proposed by the lemmatiser, but the application allows also to easily locate "unallowed lemmas" that do not exist in the reference list. When suitable lemmas do not exist in the TL, we create and add new lemmas. These are mostly cases of:

- Proper names (first name, surname, toponym).
- Words not listed in the TL's dictionary: e.g. *departement* 'departure'; *enterinement2* 'completely'.
- Lemmas listed in the TL, but reported at the end of the article: ex. adverbs (*principaument*, *mëismement*, *ancïenement*, *covenablement*); participles as nouns (*sëu*: NOMcom, TL, s. v. *savoir*), etc.

2.2.2 Tagging principles

POS Tagging is done following the rules given in the retained annotation scheme, `Cattex2009` [Guillot et al., 2013a]. Yet, some new rules had to be established and some adaptations were

made, the latter especially since we are the first to use – to our knowledge – the full annotation scheme, including morphological tags (`Cattex2009-full`).

Contraction. In the case of contractions, we have introduced dual labels; e.g., "aux" or "auquel" will be treated under "a3+le" "a3+lequel", eg., for *al*, *au*, *as* ('at the')

Agglutination/deglutination. The division of the lexical units often differs according to the date of the text and the choice of the editor, which does not necessarily match the spacing of the manuscript; one can suppose that the script of ancient manuscripts of Old French texts is roughly more analytic than those of Middle French manuscripts. Cases include, *lors que/lorsque*; *toutes voies/toutesvoies*; *par mi/parmi*; *a fin/affin*; *ja mais/jamais*, etc. Solutions should be sought that do not unduly complicate the processing of the texts. In case of agglutination or deglutination, we choose a solution adapted to the form of the text. For instance, in "Le dit jour", if "le dit" is noted in 2 words, it will be treated as `DETdef+VERppe`, whereas if it is noted in one word, *ledit* will be tagged as a compound determinant (`DETcom`). However, in accordance with the lemmatisation of Tobler and Lommatzsch [2002], complex adverbial formations such as *parfois*, *portant*, *porce*, *derechief*, *maintefois*, *naguère*, *jamais* etc. and adverbial variants with intensive prefix *tres* (`ADVgen`) are always treated analytically.

Named entities. The processing of named entities is left to an ulterior stage of corpus annotation. Regarding lemmatisation, each word is tagged according to the part of the original speech, hence the fact that "saint", part of church names or of a toponym is treated as an adjective, e.g. "dou dit priorei de *Saint Jaike on Mont*",

Saint	saint	`ADJqua`	`NOMB.=s|GENRE=m|CAS=r|DEGRE=p`
Jaike	Jacques	`NOMpro`	`NOMB.=s|GENRE=m|CAS=r`
on	en1+le	`PRE.DETdef`	`MORPH=empty+NOMB.=s|GENRE=m|CAS=r`
Mont	mont	`NOMcom`	`NOMB.=s|GENRE=m|CAS=r`

Homonyms. some very frequent homonyms deserve special attention when correcting POS and morphological tagging:

a *a3* (PRE), *avoir1* (VERcjg), *a2* (INJ);
le, la, les *il* (PROper), *le* (DETdef);
se, s' *soi1* (PROper), *ce1* (PROind), *se* (CONsub), *si* (ADVgen), *son4* (DETpos), *ce2* (DETdem);
ou, o, u *o3* (CONcoo), *où* (PROrel), *en1+le* (PRE.DETdef);
en *en1* (PRE), *en2* (PROadv), *on* (PROind);
ne *ne1* (ADVneg), *ne2* (CONcoo).

To this, one can add other homographs whose treatments requires a careful morpho-syntactic analysis and is tributary to the entries of the chosen lemma reference list and POS tagset [Tobler and Lommatzsch, 2002, Guillot et al., 2013a]. In particular, *que* can be

que1 (< QUAM) CONsub (in a comparison) or ADVgen (in constructions, *ne... que, ne mais que...*;
que2 (< QUI, QUEM, QUAM, QUOD) PROrel;
que3 (< QUID) ADVint (interrogative, exclamative);
que4 (< QUIA) CONsub (eg. "il avint que") or CONcoo ("Si sont pres de trente mile, que chevaliers que sergenz que borjois").

Regarding flexion tags, there are 12 possible composite labels for all these *que*. Yet its complexity is far from matching verbal forms, with about 60 possibilities (in mode, tense, person, number).

Irregularities, scribal or editorial mistakes, rare forms. When a rare or apparently irregular or mistaken form is encountered, the tagging becomes more difficult. For instance, we encounter phenomena of mismatches (in number, gender, verbal time) or case confusion, not only in versified texts (poetic license) but also in prose texts or practical documents. In this case, by examining the context, we choose the tagging that seems most consistent with the meaning:

- *Par Mahomet merveilles* (fem. pl.) *ai oie* (-> fem. pl. despite the form).
- *c'est assavoir de prés, de terres que* (subj. case fem. 'qui') *movoiet* (impft pers. 6) *de moi et de mon fié, que li dis abbes avoit encloses et covert* (-> fem. pl.) en son estant dou dit priorei

Diatopic variation can also cause tagging difficulties, as regional forms create new homographs. For instance, the Picard form of the singular feminine definite article or personal pronoun, *le*, homograph to the masculine forms; the possessive *se* (e.g. "*Chançon legiere a entendre Ferai,/car bien m' est mestiers /Ke chascuns le puist aprendre / Et c'on le chant volentiers*"). When we cannot decide the gender of a word, we use the value `x` especially for nouns that can be masculine or feminine, such as *ost*, *onor*, *amour*, etc.

Pyrrha's interface [Clérice and Pilla, 2021] also allows to check directly the "unallowed" values of POS tags and, more importantly, morph (because of the way the lemmatiser is trained – see below –) POS tags unseen in the training set can not be predicted, while morph tags are predicted independently (i.e., genre, numb., person, case...) and then concatenated, resulting occasionally in impossible values (e.g., comparative adjective 2nd person). In all cases, linear verification is essential for lemma disambiguation and morphosyntactic analysis. Finally, to control and homogenise the tagging result, we can revise the work by concordance on the platform itself.

The current state of the system allows a fairly easy exploitation of the corpus and the tagger begins to function properly. However, it is essential to annotate the maximum number of texts now so that the tagger learns to better discern the syntax of the texts and so that we can establish its final version in an optimal form.

III TRAINING SETUP

In order to get training samples whose structure mimics that of data observed in the real world, we ensured that our data is segmented by sentence (finishing by a `PUNfrt`) or by line (in manuscripts without punctuation). To do so, we transform the dataset using Protogenie [Clérice, 2020] which handles some form of normalisation: we normalise Roman numerals into Arabic numerals to reduce the complexity of the data, as well as the number of allowed numbers and we split the complex morphology into several simple categories (Case, Tense, etc.). The full corpus is then split into 3 different parts, for training, development and testing, with a 80/10/10% ratio.

Each task is trained separately, with a fork of Pie[Manjavacas et al., 2019], PaPie, which provides a few additions including:

- Ability to focus on various metric for improvement tracking (Accuracy, F1, Precision, Recall);
- New optimisers (Ranger for example) and learning rate schedulers;
- "Noise introduction strategies", such as randomised full capitalisation of sentences.

For each task (Lemma, POS and morphology tags), we conducted a parameter research, using prior knowledge from other languages, for which we will state the value below. Each training phase was run at least 5 times to account for random local minima and each result was logged. Once all models were trained, we implemented a ranking method that ranks each model on each

Task	Character Layers	Embedding Size	Hidden Size	Target
Lemma	2	**300**	*150*	Precision
POS	2	200	**350**	Accuracy
CAS	2	150	*150*	Accuracy
DEGRE	2	200	250	Accuracy
GENRE	2	200	250	Accuracy
MODE	2	*150*	200	Accuracy
NOMB	1	200	250	Accuracy
PERS	2	*150*	**350**	Accuracy
TEMPS	2	200	**350**	Accuracy

Table 3 – Best parameters found after the sweep. Bold are the highest value, italic the lowest.

available and meaningful metric and chose the one whose sums of ranks is the lowest. For this, we excluded metrics such as the one applied to "unknown targets" based on a very small sample of lemma which would rapidly skew the overall ranking of models.

All models shared most of the same parameters: they used a single linear layer, LSTM cells for the hidden network, a character embedding using RNN, a dropout of 0.32, learning rate of 0.0049, patience for the learning rate evolution of 2, a patience for early stopping of 5. We provide the configuration on our repository. We used the Ranger optimizer that has shown less variation in training with better scores[5].

Four parameters were used to generate diverse combinations:

- The Character Embedding size, `cemb_size`, could be of 100, 150, 200 or 300
- The number of layers for the Character Embeddings Encoder, `cemb_layers`, could be 1 or 2.
- The size of the hidden layer, that encodes most of the context, was a value in the set $150, 200, 250, 300, 350$. We added a variation only for the lemma task at 170.

In general, while the hidden size has a low impact on the lemma task, morphological tasks were more inclined to get better results with it. The number of layers always yield better except for POS, and Nomb, indicating that the task was simpler to solve for the network at the character level, most information coming from the context. Targeting precision was efficient only for the lemmatisation task. The final best parameters are shown in table 3.

IV RESULTS

4.1 Scores

Scores are shown in Table 4. They reach a global level of accuracy that is similar regarding lemma and POS (c. 97.5%), and more heterogeneous for morphological tags, from 95.25% for case to 99.05% for verbal mode. The accuracies are relatively robust to ambiguous tokens, at least for lemma and POS.

Thanks to its character level modelling, the model also achieve a **69% accuracy for lemma prediction on tokens never seen during training**.

It is to be noted that, behind the composite figure of 97.66% accuracy for lemmas, variation are to be found between grammatical categories, with 100% accuracy being reached for unambiguous and easy to tag punctuation signs or possessive adjectives (*mien*, *nostre*, etc.) and relatively low

[5]On this topic, see the discussion between *TC* and Enrique Manjavacas at https://web.archive.org/web/20210914113014/https://github.com/emanjavacas/pie/issues/76.

	All			Known tokens			Unknown tokens			Ambiguous tokens		
task	acc	pre	rec	acc	pre	rec	acc	pre	rec	acc	pre	rec
lemma	97.66	76.08	75.86	98.38	91.72	91.71	69.01	48.86	48.72	97.79	78.24	78.11
POS	97.55	83.00	79.90	97.88	84.81	82.11	84.56	54.19	55.92	97.37	83.70	81.20
CAS	95.25	91.75	92.26	95.74	92.18	92.74	84.63	62.60	62.55	94.01	91.77	92.33
DEGRE	98.47	86.45	82.77	98.74	87.75	84.98	92.55	63.81	44.48	95.46	86.91	85.80
GENRE	96.21	93.22	89.03	96.76	93.67	89.66	84.47	88.01	69.87	94.59	92.11	87.90
MODE	99.05	91.59	89.00	99.36	93.56	92.06	92.39	79.46	74.89	97.61	88.89	92.14
NOMB	97.05	96.50	96.20	97.47	96.92	96.65	88.20	87.85	82.20	96.06	95.67	95.36
PERS	98.79	91.25	85.28	98.95	91.73	85.73	95.34	90.17	90.12	97.72	91.31	85.12
TEMPS	99.13	96.51	96.20	99.38	97.21	97.58	93.94	88.92	85.39	97.80	95.02	96.78

Table 4 – Lemmatisation and tagging accuracies on the test set for the best model for each configuration. "Unknown tokens" are tokens never seen during training, while "ambiguous tokens" are forms that can correspond to different lemmas. Results for "Unknown targets" (lemmas never seen in training, but that the neural network can still accurately predict thanks to its character level modelling) are not given because the support (258 tokens) was too low for it to be significant.

scores for misspelled words (`OUT`), interjections, categories with a grammatical ambiguity or proper names (see in appendix B for detailed scores).

Some lower scores are affected by grammatical ambiguity and arbitrary lemmatisation choices. Apart from the different forms of *que*, that we will discuss in next subsection, the adjectival forms of the verbs and the qualifying adjectives have score that are driven down by this. For instance, for `VERppa`, the choice of tags for the forms in *-ant* (*vaillant*, *chantant*, *vivant*,...), that in Modern French, can be depending on the case, either adjectives or present participles, but are not so strictly classified in Old French, necessitate a more theoretical and arbitrary decision than some other categories. They are here tagged (according to `Cattex2009` principles, Guillot et al. [2013b]) always as verbal forms (`VERppa`) with a human annotated verbal infinitive as lemma (*vaillant*→*valoir*, *chantant*→*chanter*, etc.), yet the tagger is on occasion inclined to treat them as adjectives (indeed close to their actual syntactic function).

4.2 Errors and Most frequent confusions

To go beyond the constructed numeric values of scores, it is possible to inspect the confusion matrix indicating the most frequent types of wrong predictions and a classical tool in machine learning (Tables 5 and 6).

Concerning lemmas, two major conclusions can be drawn:

1. the most frequent errors are massively related to homographs, and particularly homographs of function words (*que1-4*, *ne1-ne2*, *le* as pronoun or definite determiner, etc.), that also cause trouble to human annotators (see above, subsection 2.2). Particularly, confusions between *que1*, *que2*, *que3* and *que4* account for 15.25% of all errors.
2. the most frequent errors are on function words, which is easily explainable given that they form a large majority of the total number of words for a small number of lemmas,
3. meanwhile, **around 40% of errors are distributed between 679 different lemma with a single error** (e.g., *avoutire1*, *hustin*, *Mordred*, ...), mostly proper names, content words and rare forms (which are interesting to lexicographers),
4. and a majority of errors themselves (e.g., *il* instead of *avoir*, *venir* instead of *aler*, etc.) are in single occurrence (53%), which complicates the task of batch correction.

Concerning part-of-speech tags (see detailed scores per category in appendix, Table A), a few categories achieve a perfect score due to their mostly unambiguous nature and very low diversity,

GT	Errors	Preds	Freq
que2	147	que4	137
que4	50	que2	41
que1	49	que4	46
il	42	le	36
le	26	il	24
a3	21	avoir	19
ne1	20	ne2	20
avoir	19	a3	13
que3	16	que4	10
ne2	15	ne1	15
si	14	se	9
se	13	soi1	6
on	13	en3	5
en2	12	en1	6
ce2	11	ce1	6

Table 5 – Confusion matrix for lemmas (only values > 10 were kept).

GT	Errors	Preds	Freq	GT	Errors	Preds	Freq
NOMcom	203	ADJqua	53	OUT	70	VERcjg	13
		VERppe	38	PROper	69	DETdef	22
		VERcjg	31			PROimp	17
		ADVgen	25			DETpos	15
		NOMpro	19	PROind	64	DETind	22
		VERinf	15			ADVgen	18
ADVgen	173	NOMcom	35	NOMpro	48	NOMcom	20
		CONsub	26	CONcoo	45	ADVneg	26
		PRE	21	DETdef	36	PROper	31
		PROind	16	DETcar	28	DETndf	12
		ADJqua	15			ADJcar	10
		PROper	13	ADVneg	25	CONcoo	14
VERcjg	155	VERppe	59	ADJpos	25	DETpos	18
		NOMcom	44	DETind	23	PROind	8
		PRE	21	VERinf	22	NOMcom	15
ADJqua	119	NOMcom	66	DETpos	22	PROper	5
		VERppe	23	PROcar	21	DETcar	6
CONsub	119	PROrel	83	ADJcar	20	DETcar	10
		ADVgen	24	ADJind	19	DETind	13
PROrel	98	CONsub	86	VERppa	17	ADJqua	6
PRE	78	ADVgen	37	ADVint	17	CONsub	9
		VERcjg	15	PROadv	15	PRE	6
		PROadv	10	PROint	14	PROrel	9
VERppe	78	VERcjg	38	PROdem	13	DETdem	9
		NOMcom	16	PROord	12	ADJord	8
		ADJqua	14	ADVsub	11	CONsub	4
PROimp	78	PROper	78	DETdem	10	PROdem	6

Table 6 – Confusion matrix for parts-of-speech (only values ≥ 10 were kept).

such as punctuation signs (`PON...`), while, at the other end of the spectrum the lowest scores are reached for rare and ambiguous categories like the homographic impersonal pronoun *il* (`PROimp`, vs *il* `PROper`), the homographic *que3*, *come1* and other `ADVint`, or the category `OUT`, that was left in the test set, but whose presence is arguable, because this category concerns words that are taken out of grammatical analysis (mostly scribal mistakes, for instance, such as words repeated a second time or left unfinished by the scribe). It is to be noted that these low F1

scores[6] are mostly driven down by a very low recall. It can be interpreted as a large tendency of the tagger to give to occurrences of this rarer categories for a given form or lemma the tags from one more often encountered for them in the training material, as can also be seen from the confusion matrix (Table 6).

Content-words categories (such as `NOM...`, `ADJ...` or `VER...`) achieve relatively high scores, yet their high frequency makes them featuring prominently in the confusion matrix, with some of the most frequent confusions being between common nouns and adjectives, or between the different verbal subcategories.

4.3 Qualitative inspection of lemmatisation results and comparisons with other models

For a more direct inspection by the human, we evaluate the results of the lemmatiser on a test set, which contains 2 849 random sentences drawn from the corpora. The evaluation is made on the comparison between the prediction of the model and the manually corrected gold standard. We then compute a sentence-level word-based accuracy (number of words with at least one error divided by total number of words). The distribution of sentences scores is presented in Table 7.

Sentence score	Number of sentences
1	1 871
0.9 – 1	842
0.8 – 0.9	114
< 0.8	22

Table 7 – Number of sentences for each score

The results show that most sentences are correct (around 95%). The sentence with the worst score (0.32) is actually not written in Old French but in Latin. Two other sentences have a very weak score, of 0.4 (Table 8; from now on, errors are shown in italics in the tables).

tokens	Qui	montagu	auoit	a	iustisier
correct	qui	Montagu	*aler*	a3	justicier2
predicted	qui	*montaignor*	avoir	a3	*vistoier*
TT-TL	*cuidier1*	*montagu*	avoir	*a3|a*	*justicier2|justicier*
TT-MF	qui	*montagu*	avoir	a	*iustisier*
tokens	A	Gironuille	uont	ludie	veir
correct	a3	Gironville	aler	Ludie	vëoir
predicted	a3	*Gironle*	*avoir*	Ludie	*vair1*
TT-TL	*avoir*	*Gironuille*	aler	*ludie*	*vair1|voir*
TT-DMF	a	*Gironuille*	aller	*ludie*	*voir*

Table 8 – Sentences with worst scores in the test set; row gives the original tokens, the human annotated ground truth (correct), our model prediction (predicted), and, for comparison, the results of two sets of parameters for the TreeTagger lemmatiser.

Two parameters explain the weakness of the score: first, the sentences are short, so one error strongly affects the score; secondly, these sentences present peculiar spellings, contrary to the majority of the tokens in the corpus, either because of diatopic variation or editorial choices (e.g., distinction *i/j* and *u/v*). In this particular example, we note several difficulties. One concerns the proper nouns, with *montagu* interpreted as *montaignor* and *Gironuille* as *Gironle*. An other one is the problem of multiple spellings, as *veir* as an occurrence of *vëoir* is difficult to identify. We also note that the human corrected value can sometimes be the wrong one, as *auoit* was wrongly

[6]The F1 score is the harmonic mean of precision and recall, and as such a global measure of accuracy.

identified by the human annotator as an occurrence of *aler*. The lemmatiser, here, is right. This kind of error can happen elsewhere (Table 9).

tokens	Et	bien	sachiez	qu’	en	l’	eglisse
correct	et	bien1	*sachier2*	que4	en1	le	eglise
predicted	et	bien1	savoir	que4	en1	le	eglise

Table 9 – lemmatiser can perform better than the human annotator.

This kind of error tends to show that the score that we present is at times lower than it should be. Moreover, other frequent errors which bring down the sentence scores happen because of annotators tag choice regarding capitalisation, numerals, or are due to the tags of the interjections, all that may be at times inconsistently tagged by human annotators (Table 10).

tokens	sarrasins	dex	.l.m.	O	Ha
correct	sarrasin	dieu	50000	o2	*a2*
predicted	Sarrasin	Dieu	*50*	*ho!*	ha!

Table 10 – Problems with the choice of the tag

All of these are marginal errors, that could be resolved with some additional work on the harmonisation of the corpora. Some errors are more important. They are caused by homography and often concern personal pronouns and possessive, as we can see in Table 11.

tokens	Et	veez	la	la	.
correct	et	vëoir	il	là	.
predicted	et	vëoir	*là*	là	.
TT-TL	et	*vëer|vëoir*	*le*	*là|il*	.
TT-DMF	et	vëoir	*là*	là	.

tokens	et	nos	partimes
correct	et	nos1	partir
predicted	et	*nostre*	partir
TT-TL	et	*nos|nos1|noz1*	partir
TT-DMF	et	nos	partir

tokens	li	reis	garsie	est	mis	germeins	cusins	mis	uncle	fu	fernagu
correct	le	roi	Garsie	estre1	mon1	germain	cosin	mon1	oncle	estre1	Fernagu
predicted	le	roi	Garsie	estre1	*metre2*	germain	cosin	mon1	oncle	estre1	Fernagu
TT-TL	le	roi	*garsie*	*ester| estre1*	*manoir| metre| mettre*	<*nolem*>	cosin	mon1	oncle	*estre1|estre*	*fernagu*
TT-DMF	le	roi	Garsie	estre1	*metre2*	*germeins*	cosin	mon1	oncle	estre1	Fernagu

Table 11 – Problems with personal pronouns and possessives

The errors are understandable: the form *la* can indeed refer to the adverb *là*, in the first sentence, as the form *nos*, in the second one, can be an occurrence of the personal pronoun or of a possessive. Interesting is the form *mis*, which can be an occurrence of *metre2*, which is once predicted wrong, and once right, in the same sentence.

Other errors which are caused by the existence of similar spellings can be more important (Table 12). Here, the verb *descirier*, “to tear”, is predicted as *desirrier2*, a substantive meaning “whish”. This error is important because it changes the semantics of the whole sentence.

We can compare the results of the lemmatiser with the ones of an other one, TreeTagger, for which at least two sets of parameters for Old French exist, using the lemmas from TL [Stein, s.d.],

tokens	Et	tant	mantel	desrompre	et	dessirier
correct	et	tant	mantel	desrompre	et	descirier
predicted	et	tant	mantel	desrompre	et	*desirrier2*
TT-TL	*avoir*	tant	mantel	desrompre	et	descirier
TT-DMF	et	tant	manteau	_	et	déchirer

Table 12 – Wrong prediction due to homography

henceforth **TT-TL** or from the DMF [Base de français médiéval, s.d.], henceforth **TT-DMF** [7]. The results for the two sentences with low scores tagged with TreeTagger are shown in Table 8.

The lemmatiser has the same difficulties which were mentioned above: proper nouns and particular spellings (*veir*). If its proposition for the form *iustisier* is better than the one from our model, it makes mistakes for easier occurrences (*Qui*, relative pronoun, is identified as an occurrence of *cuidier1*, "to believe"; *A*, preposition, is identified as an occurrence of the verb *avoir*). It also encounters difficulties with personal pronouns and possessive and produces the same errors as we saw in table 11.

The TT-TL and TT-DMF parameters produce one error our model didn't produce: it cannot identify the form *germeins*. The TT-TL parameters, on the test set, provides for 1,561 occurrences a <*nolem*> tag, and the TT-DMF ones, a "_" tag, as in table 12, for 2,878 occurrences. The TT-TL parameters also propose alternative choices where our model proposes just one.

On occasions, our model is seen better handling spellings marked from a diachronic or diatopic perspective, or more generally, less familiar spellings. For instance, it handles better the Anglo-Norman forms seen in 13, where TT-TL makes six mistakes, TT-DMF makes five and our model none (but it is a sample drawn from an in-domain text).

tokens	laisum	clarel	cest	saracin	aler	,	kar	bin	vez	nel	pouum	mener
correct	laissier	Clarel	cest	Sarrasin	aler	,	car	bien1	vëoir	ne1+il	pöoir	mener
predicted	laissier	Clarel	cest	Sarrasin	aler	,	car	bien1	vëoir	ne1+il	pöoir	mener
TT-TL	laissier	*clarel*	cest	saracin	*aler|foraler*	,	car	<*nolem*>	*aler|vëoir*	*il|ne1*	pouvoir	mener
TT-DMF	laisser	*clarel*	cest	saracin	aller	,	car	*bin*	*fois*	ne.il	*pouum*	mener

Table 13 – Comparisons of the results on one sentence with Anglo-Norman forms

It can happen that the results of TreeTagger are better than ours, for instance, in the sentence with the difficult form *dessirier* (Table 12). However, TT-TL makes a mistake for the identification of the easy form *Et*, which cannot be identified, perhaps due to capitalisation. Moreover, it on turns produces hesitations or mistakes on similar forms (Table 14).

tokens	As	paiens	sont	venu	,	de	ferir	desirant
correct	a3+le	paien	estre1	venir	,	de	ferir	desirrer
predicted	a3+le	paien	estre1	venir	,	de	ferir	desirrer
TT-TL	*a+le|le*	paien	*estre1|estre*	venir	,	de	ferir	*descirier|desirer*
TT-DMF	a.le	païen	être	venir	,	de	férir	*déchirer*

Table 14 – Prediction of TreeTagger on the form *desirant*

[7] They are available at: https://sites.google.com/site/achimstein/research/resources and https://www.cis.lmu.de/~schmid/tools/TreeTagger/.

V DISCUSSION AND FURTHER RESEARCH

Current models display very satisfying results, in particular when accounting for the specific difficulties of an under-resourced variation-rich non-standardised historical language, in which variation also increases the number of homograph forms and the size of the "vocabulary" of forms. The use of a neural tagger and the creation of a significant gold corpus has allowed to obtain very usable results, especially when tested on in-domain material.

In some cases, it might be possible to achieve further gains in accuracy by crossing predictions for lemma and for part-of-speech (which are for now predicted independently by the neural network). This could be the case, for instance, for the various occurrences of *que<n>* (Table 15).

GT	Freq	Acc. POS	Acc. Lemma	Combined Accs	Pred=*que4*	Pred=*que2*
que1	97	**82.47**	48.45	39.18	48.45	3.09
que2	587	**84.67**	75.98	72.57	22.49	75.98
que3	37	**67.57**	59.46	56.76	24.32	16.22
que4	1474	92.13	**96.40**	90.84	96.40	2.92

Table 15 – Lemma and POS accuracy of the prediction for the different kinds of *que*; better performance of the POS indicates that in many cases lemmatisation accuracy could be increased through post-treatment after the lemma prediction phase.

Nonetheless, the generality of the models will need to be evaluated in a more systematic fashion, regarding the variety of regional scriptae and written genres of Old French. For this, we will need to build an out-of-domain corpus to get a more general evaluation of the model, and to test its performances on specific scriptae or genre (What are the performances on Picard or Lorrain texts? On theatre? etc.).

Another interesting lead is to explore transfer learning approaches on neighbouring languages, such as Occitan for which some early experiments have been conducted [Camps and Couffignal, 2017], on later varieties such as Middle and Pre-Classical French, and on Franco-Romance hybrids, in particular Franco-Italian, for which the creation of an annotated corpus is ongoing [Gambino, Ceresato, to be published].

Diachrony should also be explored. For the moment, For the moment the corpus includes only one text from the 14th century, an extract from the *Pèlerinage de l'âme* by Guillaume de Digulleville (1355-1358). Given the costs of building lemmatised corpora, one must be careful not to create duplicates. For this reason, an extension towards Middle French would need to be done in order to maintain interoperability of the corpora with the *Dictionnaire de Moyen Français* and the tools developed around it. Integrating this period would fill the current break between the models we make available for Old French and for Early Modern French [Camps et al., 2020] in order to allow for studies in long diachrony [ATILF, 2015].

More generally, further work needs to investigate the interoperability of annotation schemes. One possibility could be to offer a version of the corpora converted to the use of Universal POS tags (UD POS), that are of broad use today for many Contemporary languages. In any case, there is a need for the elaboration of a common canonical reference tag-set and lemma list, to which and from which designing conversions for all the formats that are currently being used by the projects working on Old French lemmatisation.

AUTHORS CONTRIBUTIONS

JBC started and over-viewed the projects, trained some of the models, took part alone or with others in the annotation or correction of the corpora Geste, worked on the workflows and reference lists; he coordinated the team, with FD. **TC** worked mostly on the engineering and NLP research around this corpus, built the interfaces that are used (Pyrrha, Deucalion), organized the training and optimization of the models, expanded tools to ensure the applicability outside of tests (PaPie) and engineered the various tools that ensure reproducibility and quality control (Protogenie, PyrrhaCI). He provided some insights when necessary regarding some annotation choices and technological limitations. **FD** worked on the post-correction of juridical texts and coordinated the team with JBC. **LI** joined the project during its early phase and worked on the correction of the texts in the LAKME and OMÉLiE projects; she also provided data from her PhD (in progress). **NK** worked on the post-correction (lemma, POS and morphosyntactic labelling) of texts that were used to improve automated tagging performance. **AP** joined JBC in annotating data in parallel of the LAKME project with data from her PhD Thesis, provided most of the use-cases and early tests of Pyrrha by correcting her corpus (lemma and morpho-syntax).

The authors have no competing interests to declare.

MATERIALS AND DATA AVAILABILITY

The most up-to-date version of the models can be easily obtained and used thanks to the `pie-extended` Python package, available on Pypi (https://pypi.org/project/pie-extended/), with the command `pie-extended download fr`, and can be queried at https://tal.chartes.psl.eu/deucalion/.

ACKNOWLEDGEMENTS

This work benefited from several projects funded by various institutions: LAKME – *Linguistically Annotated Corpora Using Machine Learning Techniques* (2016-2018, Université PSL); OMÉLiE – *Outils et méthodes pour l'édition linguistique enrichie* (2018-…, Université PSL/IRIS Scripta and Région Île-de-France, DIM *Sciences du texte et connaissances nouvelles*), as well as, for the legal corpus, of *Biblissima*. We thank the DIM *Science du texte et connaissances nouvelles* for funding the acquisition of a GPU server, as well as the École nationale des chartes for providing infrastructure and support for the server.

We thank Mike Kestemont for the collaboration regarding neural lemmatisers, as well as Enrique Manjavacas for his precious advice regarding lemmatisation and Pie configuration. We also thank Thierry Poibeau (LATTICE) and Daniel Stoekl (EPHE), with which the initial LAKME project was launched. Our gratitude for the collaboration also goes to the team working on Franco-Italian texts, led by Francesca Gambino. It is difficult to acknowledge the help of all colleagues that helped along the years, but we non exhaustively thank Floriana Ceresato, Simon Gabay, Sophie Prevost as well as all the participants in the workshop "Référentiels de lemmes du français médiéval", in particular, Alexei Lavrentiev, Martin Glessgen, Simon Gaunt, Maud Becker and Gilles Souvay, without forgetting the students of the École des chartes.

References

ATILF. *Dictionnaire du Moyen Français (DMF 2015)*. CNRS and Université de Lorraine, Nancy, 2015. URL http://www.atilf.fr/dmf.

Base de français médiéval. TreeTagger: Parameters for Old French, s.d. URL https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/old-french.par.gz.

Gace Brulé and Holger Niels Dyggve Petersen. *Gace Brulé, trouvère champenois: édition des chansons et étude historique*. Number 16 in Mémoires de la Société néophilologique de Helsinki. Société néophilologique, Helsinki, 1951.

R. Busa. The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities*, 14 (2):83–90, 1980. ISSN 0010-4817. URL https://www.jstor.org/stable/30207304. Publisher: Springer.

Florian Cafiero and Jean-Baptiste Camps. Why Molière most likely did write his plays. *Science Advances*, 5(11), 2019. doi: 10.1126/sciadv.aax5489. URL https://advances.sciencemag.org/content/5/11/eaax5489.

Jean-Baptiste Camps. *La 'Chanson d'Otinel': édition complète du corpus manuscrit et prolégomènes à l'édition critique*. thèse de doctorat, dir. Dominique Boutet, Paris-Sorbonne, Paris, December 2016. URL https://halshs.archives-ouvertes.fr/tel-01664932.

Jean-Baptiste Camps, editor. *Geste: un corpus de chansons de geste, 2016-... (Version 02)*. École nationale des chartes, Paris, April 2019. URL http://doi.org/10.5281/zenodo.2630574. textes du domaine public, développements CC-BY-SA.

Jean-Baptiste Camps and Gilles Guilhem Couffignal. La production de corpus d'occitan médiéval et prémoderne: problèmes et perspectives de travail. In *Fidelitats e dissidencias/Fidélités et Dissidences, Actes du XI\ieme{} Congrès International de l'Association Internatioale d'Études Occitanes, Albi 2017*, July 2017. doi: 10.5281/zenodo.2605497. URL https://halshs.archives-ouvertes.fr/halshs-02050089/. à paraître.

Jean-Baptiste Camps, Lucence Ing, and Elena Spadini. Collating Medieval Vernacular Texts. Aligning Witnesses, Classifying Variants. In *DH2019 Digital Humanities Conference 2019*, Utrecht, Netherlands, July 2019. URL https://hal.archives-ouvertes.fr/hal-02268348.

Jean-Baptiste Camps, Simon Gabay, Paul Fièvre, Thibault Clérice, and Florian Cafiero. Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre. *arXiv preprint arXiv:2005.07505*, 2020.

Floriana Ceresato. L'analisi lessicale dell'entrée d'espagne: bilancio di una prima sperimentazion. *Francigena*, to be published.

Thibault Clérice. Protogenie, post-processing for nlp dataset, June 2020. URL https://doi.org/10.5281/zenodo.3883585.

Thibault Clérice. Pie Extended, an extension for Pie with pre-processing and post-processing, 6 2021. URL https://github.com/hipster-philology/nlp-pie-taggers.

Thibault Clérice and Julien Pilla. Pyrrha, 7 2021. URL https://github.com/hipster-philology/pyrrha.

Conon de Béthune. *Les chansons de Conon de Béthune*. CFMA. Paris, 1925. URL https://gallica.bnf.fr/ark:/12148/bpt6k977f.

Anthonij Dees, Pieter van Reenen, and Johan A De Vries. *Atlas des formes et des constructions des chartes françaises du XIII^{e} siècle*. Number 178 in Beihefte zur Zeitschrift für romanische Philologie. M. Niemeyer Verlag, Tübingen, 1980. ISBN 3-484-52084-1. URL http://dx.doi.org/10.1515/9783111328980.

Anthonij Dees, Marcel Dekker, Onno Huber, and Karin Van Reenen-Stein. *Atlas des formes linguistiques des textes littéraires de l'ancien français*. De Gruyter, Berlin, Boston, reprint 2014 edition, 1987. ISBN 978-3-484-52212-1. URL https://www.degruyter.com/viewbooktoc/product/160190.

Oksana Dereza. Lemmatisation for under-resourced languages with sequence-to-sequence learning: A case of early irish. In *Proceedings of Third Workshop Computational linguistics and language science*, volume 4, pages 113–124, 2019. doi: 10.29007/cxtl.

Eglal Éditeur scientifique Doss-Quinby, Joan Tasker Éditeur scientifique Grimbert, Wendy Éditeur scientifique Pfeffer, and Elizabeth Éditeur scientifique Aubrey. *Songs of the women 'trouvères'*. New Haven, 2001. ISBN 978-0-300-08412-2.

S. Feldman, M. A. Marin, M. Ostendorf, and M. R. Gupta. Part-of-speech histograms for genre classification of text. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4781–4784, Taipei, Taiwan, April 2009. IEEE. ISBN 978-1-4244-2353-8. doi: 10.1109/ICASSP.2009.4960700. URL http://ieeexplore.ieee.org/document/4960700/.

Francesca Gambino, editor. *RIALFrI – Repertorio Informatizzato Antica Letteratura Franco-Italiana*. URL http://www.rialfri.eu/.

Martin Dietrich Gleßgen. *Les plus anciens documents linguistiques de la France*. 2016. URL http://www.rose.uzh.ch/docling/. 3^{e} édition.

Céline Guillot, Sophie Prévost, and Alexei Lavrentiev. *Manuel de référence du jeu Cattex09*. École normale supérieure de Lyon, Lyon, 2013a. URL http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf. Version 2.0 – 8 avril 2013.

Céline Guillot, Sophie Prévost, and Alexei Lavrentiev. Principes d'annotation cattex09. Technical report, École normale supérieure de Lyon, Lyon, 2013b. version 2.0. http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_principes_2.0.pdf.

Lucence Ing. *Disparitions lexicales en diachronie: traitements automatiques sur le Lancelot en prose*. Phd thesis, dir. f. duval, codir. j.b. camps, Université PSL, Paris, 2021.

Mike Kestemont and Jeroen De Gussem. Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning. *Journal of Data Mining & Digital Humanities*, August 2017. URL https://jdmdh.episciences.org/3835. arXiv: 1603.01597.

Mike Kestemont, Guy de Pauw, Renske van Nie, and Walter Daelemans. Lemmatization for variation-rich languages using deep learning. *Digital Scholarship in the Humanities*, 32:fqw034, August 2016. ISSN 2055-7671, 2055-768X. doi: 10.1093/llc/fqw034. URL http://dsh.oxfordjournals.org/content/early/2016/08/26/llc.fqw034.

Pierre Kunstmann, editor. *Chrétien de Troyes: Cligès, Erec, Lancelot, Perceval, Yvain – manuscrit P (BnF fr. 794)*. 2009. URL http://www.atilf.fr/dect.

Pierre Kunstmann and Achim Stein. *Le nouveau corpus d'Amsterdam: actes de l'atelier de Lauterbad, 23-26 février 2006*. F. Steiner, Stuttgart, Allemagne, 2007. ISBN 978-3-515-08997-5. ISSN: 0341-0811.

Enrique Manjavacas, Thibault Clérice, and Mike Kestemont. emanjavacas/pie v0.2.3, April 2019. URL https://zenodo.org/record/2654987#.XP-nvC3M0fM.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. Improving Lemmatization of Non-Standard Languages with Joint Learning. *arXiv e-prints*, art. arXiv:1903.06939, Mar 2019. URL https://www.aclweb.org/anthology/N19-1153/.

Enrique Manjavacas, Akos Kádár, and Mike Kestemont. Improving lemmatization of non-standard languages with joint learning. *arXiv preprint arXiv:1903.06939*, pages 1493–1503, June 2019. doi: 10.18653/v1/N19-1153. URL https://www.aclweb.org/anthology/N19-1153.

Sylvie Mellet. La lemmatisation et l'encodage grammatical permettent-ils de reconnaître l'auteur d'un texte? *Médiévales*, 21(42):13–26, 2002. doi: 10.3406/medi.2002.1536. URL https://www.persee.fr/doc/medi_0751-2708_2002_num_21_42_1536. Publisher: Persée - Portail des revues scientifiques en SHS.

Sylvie Mellet and Gérald Purnelle. Les atouts multiples de la lemmatisation: l'exemple du latin. In *JADT 2002, 6es Journées internationales d'Analyse statistique des Données Textuelles*, pages 529–538. INRIA, IRISA, 2002. URL https://hal.univ-cotedazur.fr/hal-01365515.

Ariane Pinche. *Édition nativement numérique du recueil hagiographique 'Li Seint Confessor' de Wauchier de Denain d'après le manuscrit fr. 412 de la Bibliothèque nationale de France*. Phd thesis, dir. c. pierreville and b. bureau, Université de Lyon, Lyon, 2021.

Mary A. Rouse and Richard H. Rouse. La concordance verbale des Écritures. In Pierre Riché, Guy Lobrichon, and Michel Zink, editors, *Le Moyen âge et la Bible*, pages 115–122. Beauchesne, Paris, France, 1984. ISBN 978-2-7010-1091-5. ISSN: 0767-0826.

Helmut Schmid. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. 1994. URL https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Helmut Schmid. Deep learning-based morphological taggers and lemmatizers for annotating historical texts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, pages 133–137, 2019.

Achim Stein. *Nouveau Corpus d'Amsterdam. Corpus informatique de textes littéraires d'ancien français (ca 1150-1350)*.

Achim Stein. TreeTagger: Parameters for Old French, s.d. URL https://sites.google.com/site/achimstein/research/resources.

Thibaud IV. *Les chansons de Thibaut de Champagne, roi de Navarre*. CFMA. 1925. URL https://gallica.bnf.fr/ark:/12148/bpt6k53236.

Adolf Tobler and Erhard Friedrich Lommatzsch. *Altfranzösisches Wörterbuch: édition électronique*. F. Steiner, Stuttgart, 2002. ISBN 978-3-515-08199-3.

A DETAILED SCORES PER PART-OF-SPEECH

target	precision	recall	f1-score	support
PONfbl	1.00	1.00	1.00	4746
PONfrt	1.00	1.00	1.00	2764
PONpdr	1.00	1.00	1.00	450
PONpga	1.00	1.00	1.00	766
PROper.PROper	1.00	1.00	1.00	13
ADVneg.PROper	1.00	0.98	0.99	54
VERcjg	0.99	0.98	0.99	9745
CONcoo	0.99	0.99	0.99	3982
DETdef	0.99	0.99	0.99	3662
PRE	0.99	0.99	0.99	5618
PRE.DETdef	0.99	0.99	0.99	979
PROdem	0.99	0.99	0.99	913
ADVneg	0.98	0.98	0.98	1543
NOMcom	0.98	0.98	0.98	10007
PROadv	0.98	0.98	0.98	811
PROper	0.98	0.99	0.98	6168
VERinf	0.98	0.99	0.98	1537
DETpos	0.97	0.98	0.98	1361
NOMpro	0.98	0.97	0.97	1715
ADVgen	0.97	0.97	0.97	4999
INJ	0.97	0.97	0.97	33
DETdem	0.96	0.98	0.97	413
DETndf	0.95	0.98	0.96	403
CONsub	0.95	0.95	0.95	2578
VERppe	0.94	0.96	0.95	2213
ADVgen.PROper	0.91	1.00	0.95	10
ADJqua	0.94	0.93	0.94	1800
PROrel	0.93	0.94	0.94	1630
DETind	0.92	0.96	0.94	591
PROind	0.93	0.91	0.92	681
PRE.PROdem	0.88	0.93	0.90	15
ADJind	0.89	0.89	0.89	171
VERppa	0.89	0.87	0.88	134
DETcar	0.87	0.84	0.85	171
PROint	0.91	0.78	0.84	65
PROpos	0.84	0.84	0.84	43
ADJord	0.76	0.92	0.83	37
ADVsub	0.85	0.79	0.82	52
PROcar	0.82	0.79	0.81	101
DETrel	0.83	0.68	0.75	22
ADJcar	0.77	0.73	0.75	75
DETint	0.71	0.77	0.74	13
ADJpos	0.84	0.62	0.71	66
PROord	1.00	0.52	0.68	25
ADVint	0.78	0.55	0.65	38
OUT	0.91	0.49	0.64	137
PROimp	0.80	0.46	0.59	145

Table 16 – Detailed scores for each part-of-speech in the test set, giving precision, recall, F1 and support. POS with total frequency < 10 were removed (PRE.DETrel, PRE.PROper, DETcom, DETord, RED).

B LEMMA SCORES FOR EACH PART-OF-SPEECH CATEGORY

POS	Lemmas Accuracy	Freq	Lemmas SDI
PONfrt	100.00	2764	0.34
PONpga	100.00	766	0.00
PROimp	100.00	145	0.00
ADJpos	100.00	66	1.87
ADVneg.PROper	100.00	54	0.00
DETrel	100.00	22	0.30
PRE.PROdem	100.00	15	0.00
DETint	100.00	13	0.00
PROper.PROper	100.00	13	0.00
ADVgen.PROper	100.00	10	0.00
PONfbl	99.96	4746	0.55
DETind	99.66	591	1.87
ADJind	99.42	171	1.28
DETdef	99.34	3662	0.01
PONpdr	99.33	450	0.00
CONcoo	99.27	3982	0.99
DETndf	99.26	403	0.00
PRE	99.07	5618	2.19
PROper	98.96	6168	1.30
DETpos	98.90	1361	1.48
PRE.DETdef	98.88	979	1.00
PROdem	98.58	913	0.81
ADVneg	98.38	1543	0.92
PROadv	98.27	811	0.65
VERcjg	97.76	9745	3.96
ADVgen	97.76	4999	3.83
VERinf	97.46	1537	4.99
ADJord	97.30	37	1.87
DETdem	97.09	413	1.16
PROind	97.06	681	2.35
NOMcom	96.80	10007	6.10
ADVsub	96.15	52	0.70
PROcar	96.04	101	2.82
ADJcar	96.00	75	2.03
CONsub	95.62	2578	1.26
VERppe	95.62	2213	5.42
ADJqua	95.56	1800	4.31
PROpos	93.02	43	1.48
PROord	92.00	25	1.91
NOMpro	91.49	1715	5.17
PROint	90.77	65	1.52
PROrel	90.18	1630	1.33
DETcar	90.06	171	3.11
VERppa	88.06	134	3.90
INJ	84.85	33	1.68
ADVint	78.95	38	1.15
OUT	62.04	137	1.40

Table 17 – Detailed scores of the lemmas for each part-of-speech, giving lemma accuracy, total frequency of the tag, and the Shannon Diversity Index (0 means no diversity). POS with total frequency < 10 were removed (PRE.DETrel, PRE.PROper, DETcom, DETord, RED).