Building an Open Morphological Lexicon and Lemmatizing Old French Texts with the TXM Platform

Alexei Lavrentiev, Serge Heiden, Matthieu Decorde

А.М. Лаврентьев, С. Эйден, М. Декорд
A. Lavrentiev, S. Heiden, M. Decorde

Создание открытого морфологического словаря и лемматизация старофранцузских текстов с использованием платформы TXM

Building an Open Morphological Lexicon and Lemmatizing Old French Texts with the TXM Platform[1]

Аннотация. В докладе представлен опыт лемматизации средневековых французских текстов (IX – XV вв.) с помощью платформы TXM. В проекте доступные лексические ресурсы были использованы для построения открытого морфологического словаря старофранцузского языка (FROLEX), который, в свою очередь, использовался для автоматической лемматизации. На финальном этапе леммы проверяются и корректируются экспертом. Предложенные методологические решения и разработанный модуль морфологических словарей на платформе TXM могут использоваться для работы с другими языками, в особенности с теми, где высока вариативность написания и сегментации слов.

Abstract. This paper presents an experience of lemmatizing Medieval French texts (9th – 15th centuries) with the TXM platform (http://textometrie.org). The project uses available lexical resources to compile an open morphological lexicon of Medieval French (FROLEX), which is used in its turn to perform automatic lemmatization. At the final stage, the lemmas are verified and corrected by a human expert. The methodological solutions proposed and the tools for managing lexicons and applying lemmatization developed for TXM may be used for processing other languages, especially those with high variation in spelling and word segmentation practices.

Keywords: lemmatization, open morphological lexicon, Old French, TXM platform

Ключевые слова: лемматизация, открытый морфологический словарь, старофранцузский язык, платформа TXM

[1]The results presented in this paper were obtained in the framework of the PaLaFra Research Project (ANR-14-FRAL-0006) financed by the French National Research Agency (ANR) and the German DFG granting agency.

The lemmatization of texts in historical language corpora where word forms vary a lot depending on chronological, dialectal and individual factors has always been a challenging task [Piotrowski 2012: 96; Glessgen 2003]. Even within a single text, the variation in spelling and word segmentation may be considerable. The choice of the authority lemma form may also be a problem, as different reference dictionaries sometimes use different entry forms for the same lexeme. For these reasons the value of quality lemmatization is particularly high for historical corpora.

As far as the French language is concerned there exists a number of digitized and natively digital dictionaries (such as Tobler-Lommatzsch[2], DMF[3], DÉCT[4] or AND[5]), as well as a few lemmatized corpora (NCA[6], DÉCT corpus). Some tools for automatic or computer assisted lemmatization of Medieval French are also available. The NCA corpus comes with a morphological lexicon (AFRLEX) where the word forms are associated with lemmas from several sources. It can be used with TreeTagger software [Schmid 1995] but the lemmas it provides are too complex to be convenient for corpus users, as in the following example:

esjoir|esjöir|jouir_*I|IdTd|MMd (1)

Here, "esjoir" comes from the LFA lexicon (Ottawa University), "esjöir" from the Tobler-Lommatzsch (TL) dictionary, and "jouir" from the verb form list compiled by Robert Martin (at early stages of the DMF project).

The LGeRM tool[7] created for the work on the DMF dictionary offers an online lemmatization service but its output requires heavy

[2]http://www.uni-stuttgart.de/lingrom/stein/tl/index.htm
[3]http://www.atilf.fr/dmf
[4]http://www.atilf.fr/dect/
[5]http://www.anglo-norman.net
[6]http://www.uni-stuttgart.de/lingrom/stein/corpus/
[7]http://www.atilf.fr/LGeRM/

human work for disambiguation, as it provides all possible lemmas for a given form regardless of context and of morphological tags. The morphological lexicon of the DMF (DMFLEX) can be downloaded from the LGeRM website.

The PALM platform[8] offers an interface for computer assisted lemmatization but does not allow customizing morphological tagsets and has limited import/export capacities.

Our aim in this project is to lemmatize the texts of the *Base de français médiéval* (BFM)[9] using an open morphological lexicon compiled from the best resources available. For the users' convenience, the lemmas should be connected where possible to the online dictionary entries. The BFM morphological tags verified by human experts in nearly 25% of the corpus can be used for the primary disambiguation.

BFM includes five lemmatized texts by Chrétien de Troyes provided by the DÉCT project. DÉCT uses TL lemmas where possible and adds some of its own. These texts were morphologically tagged using the BFM language model and verified by experts in Medieval French linguistics. They form therefore the basis of the BFM morphological lexicon (BFMLEX).

The first step was to compare the morphological tagsets of AFRLEX and DMFLEX with that of the BFM [Guillot et al. 2013], and to work out conversion rules in order to merge the lexicons. Different tagsets provide unequal level of detail in morphological description, so the joint tagset has to be less detailed than any of the initial ones. For instance, in the merger of BFM (Cattex 2009) and AFRLEX tagsets we had to erase the information on the verb form classes (finite, participle or infinitive) from the BFM tags, as in AFRLEX the verb tags are not sub-categorized. Even more information is erased in the merger with DMFLEX where the distinction between adjectives, pronouns and determiners is not made

[8] http://palm.huma-num.fr/PALM/

[9] http://txm.bfm-corpus.org

systematically. In addition, some DMFLEX lemmas have double or triple morphological tags, as in the following example:

néant subst. masc., adv. et pron. indéf. (2)

Here, the lemma *néant* is associated with three morphological categories: noun, adverb and indefinite pronoun. In these case, separate lemma entries are created in the merger for each association of lemma form with a single morphological category.

As for the form of lemmas, most of AFRLEX lemmas are taken from the TL dictionary entries [Kunstmann et al. 2007], while the DMFLEX uses, where possible, the modern French forms. The choices of lemma forms (from the variety of spellings found of the Old French texts) are not entirely homogeneous throughout the TL dictionary, due to the long history of its compilation (over 80 years) and to the absence of "standard" spelling in the Old French. The choice of modern lemma forms ensures the compatibility with modern lexicons (such as that of the TLF dictionary[10]), which is convenient for compiling large diachronical corpora. However, the words that disappeared (or became extremely rare) in the history of the French language are problematic: the DMF uses either the modernized forms that look artificial (such as *cuider* for the Old French *cuidier*, 'to think') or keeps old forms (such as *estovoir* that should have given **étovoir* if the word existed in modern French), which introduces a certain kind of heterogeneity to the lemma list. The rules of creating DMF entry forms are presented in [Martin 1998: 970-973].

In the merger of BFMLEX, AFRLEX and DMFLEX into FROLEX the lemma form of the DMFLEX was preferred, and a "lemma_src" column was created to record the information on the lemma source. A separate table was created to provide correspondences between lemma forms from different sources.

The second step was to develop an extension for the TXM platform [Heiden et al. 2010][11] for working with morphological lexicons. This extension includes commands for importing lexicons in

[10] *Trésor de la langue française*, http://www.atilf.fr/tlfi

[11] http://textometrie.org

TSV format, for querying different columns using regular expressions, sorting entries, recoding morphological tags, merging lexicons and exporting the compiled lexicon in TreeTagger format. It also includes a set of commands for operating TreeTagger from the TXM interface: train, apply, project lemmas and remove properties. This extension is already available for public beta-testing from an update site dedicated to the PALAFRA project[12].

The third step, which is currently under way, consists in developing a concordance based user interface for verifying and correcting automatically tagged lemmas of a TXM corpus. While this work is in progress, the verification of lemmas can be done in a spreadsheet software (Libre Office Calc or Miscrosoft Word) thanks to annotation concordances export and import macros.

The first version of the open Medieval French lexicon FROLEX has been published on the GitHub platform under an open-source license so that the NLP community can use it and share its enrichment[13].

References

1. *Glessgen M.* (2003), La lemmatisation de textes d'ancien français: méthodes et recherches [Lemmatization of Old French texts: methods and research]. Kunstmann P. et al. (ed.), Ancien et moyen français sur le Web. Enjeux méthodologiques et analyse du discours [Old and Middle French on the Web. Methodological issues and discourse analyses], Ottawa, Les Éditions David, pp. 55-75.

2. *Guillot C., Prévost S., Lavrentiev A.* (2013), Manuel de référence du jeu Cattex09 [Reference Manual for Cattex09 tagset], Lyon, Équipe BFM, available at: bfm.ens-lyon.fr/spip.php?article323

3. *Heiden S.* (2010), The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Nov. 2010. Sendai, Japan,

[12]http://textometrie.ens-lyon.fr/dist/palafra

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

Institute for Digital Enhancement of Cognitive Development, Waseda University, pp. 389-398, available at halshs.archives-ouvertes.fr/halshs-00549764.

4. *Kunstmann P., Stein A.* (2007), Le nouveau corpus d'Amsterdam [The new Amsterdam corpus]. Kunstmann P., Stein A. (ed.), Le nouveau corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006 [The New Amsterdam corpus. Proceeding of Lauterbad workshop, 23-26 February 2006], Stuttgart: Steiner, pp. 9-27.

5. *Martin R.* (1998), Le Dictionnaire du moyen français (DMF) [The Dictionary of Middle French], Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres [Reports of Assemblies of the Academie des Inscriptions et Belles-Lettres], Vol. 142(4), pp. 961-982, available at www.persee.fr.

6. *Piotrowski M.* (2012), Natural language processing for historical texts, Morgan & Claypool Publishers.

7. *Schmid H.* (1995), Improvements in part-of-speech tagging with an application to German. Proceedings of the ACL SIGDAT-Workshop, Dublin, pp. 47–50.

Lavrentiev Alexei
Centre national de la recherche scientifique (France).
E-mail: alexei.lavrentev@ens-lyon.fr

Heiden Serge
École normale supérieure de Lyon (France).
E-mail: slh@ens-lyon.fr

Decorde Matthieu
École normale supérieure de Lyon (France).
E-mail: matthieu.decorde@ens-lyon.fr