Challenges of language change and variation: towards an extended treebank of Medieval French

Mathilde Regnault, Sophie Prévost, Éric Villemonte de la Clergerie

Challenges of language change and variation: towards an extended treebank of Medieval French

Mathilde Regnault[1,2] and **Sophie Prévost**[1] and **Eric Villemonte de la Clergerie**[2]
mathilde.regnault@sorbonne-nouvelle.fr
sophie.prevost@ens.fr
Eric.De_La_Clergerie@inria.fr
(1) Lattice, 1 rue Maurice Arnoux, 92120 Montrouge, France
(2) Inria, 2 rue Simone Iff, 75012 Paris, France

Abstract

In order to automatically extend a treebank of Old French (9th-13th c.) with new texts in Old and Middle French (14th-15th c.), we need to adapt tools for syntactic annotation. However, these stages of French are subjected to great variation, and parsing historical texts remains an issue. We chose to adapt a symbolic system, the French Metagrammar (FRMG), and develop a lexicon comparable to the Lefff lexicon for Old and Middle French. The final goal of our project is to model the evolution of language through the whole period of Medieval French (9th-15th c.).

1 Introduction

With the rise of digital humanities, more and more ancient texts are made available. Annotating them and keeping this information in treebanks helps study and describe old stages of a language. Some are available for Medieval French (9th-15th c.), namely the MCVF[1] (Martineau, 2008), annotated with constituency syntax, and the SRCMF[2] (Prévost and Stein, 2013), annotated with dependency syntax and covering Old French (9th-13th c.) for now. Our goal is to automatically extend the SRCMF treebank to obtain a larger resource. In particular, we want to add texts of Middle French, the next stage in the evolution of French (14th-15th c.), as well as new texts of Old French. This new resource would then contain one million words, four times more than the current SRCMF. We want to annotate these data automatically with the highest quality, which means we need to find a way to parse both Old and Middle French. However, this task is difficult because we have limited resources annotated with dependency syntax in Old French, and none in Middle French (Guibon et al., 2014). Moreover, Medieval French, like Old French, is subjected to great variation (sections 2 and 3).

The new texts will be annotated by both a statistical and a symbolic parser. The annotation will then be merged to obtain the best possible analysis. For this work, we focus on the symbolic approach. Using wide coverage grammars (Oepen et al., 2004; Rocio et al., 2003; Brants et al., 2002) has shown effective (section 4), so we chose to adapt the French Metagrammar (FRMG, Villemonte de la Clergerie (2005; 2013)), a symbolic system for contemporary French (section 5). Our contribution is to make a diachronic grammar for Medieval French, currently still in process.

2 Parsing historical texts

The extension of the SRCMF treebank will contain not only new texts in Middle French, but also in Old French, which will give a more accurate view of the period. Six dialects from the northern half of the territory will be added, and the representation of the different domains and genres will be better balanced owing to the new texts. For example, there is only one historical work in the latest version of the SRCMF, which prevents from drawing thorough comparisons with other domains.

However, the heterogeneity of these data is challenging. An automatic system is very unlikely to give the correct analysis of phenomena that had few occurrences (or none) in the training data. A grammar is

[1]The MCVF treebank (*Modéliser le changement : les voies du français*) is available at this address: `http://www.voies.uottawa.ca/`.

[2]The SRCMF treebank (Syntactic Reference Corpus of Medieval French) is available at this address: `http://srcmf.org/`.

also subjected to such limitation because developers rely on existing descriptions and treebanks. Moreover, the amount of available data annotated with dependency syntax is limited, and only exists for Old French. We considered several methods to solve this parsing challenge. Guibon et al. (2014) investigated statistical parsing of Old French on the SRCMF with the *Mate* parser (Bohnet, 2010). They obtained an average labelled attachment score (LAS) of 76.04. They developed a methodology relying on the selection of a training set with metadata[3] similar to the new text to annotate with a minimum error rate (Guibon et al., 2015).

Although it would be possible to use such a method to extend the SRCMF treebank, statistical parsing still heavily depends on the training data. The solution we chose is to adapt a symbolic or hybrid system built for French, following Rocio et al. (2003) for Old Portuguese. However, that system will have to be flexible enough to enable the treatment of the great variability of Medieval French.

3 Difficulties due to the specificities of Medieval French

Even though Contemporary French largely differs from Medieval French, there are still enough similarities to enable us to adapt a grammar. Almost all syntactic phenomena in French are already present in Medieval French, but with different frequencies. For example, SVO became the prevalent word-order as early as the 11th century (it was previously SOV, inherited from Latin, which was prevalent). In case of complete absence of information on words, it is possible to resort to a morphological and syntactic lexicon of contemporary French. The descriptions of syntactic phenomena should however be modified to parse Medieval French and cope with linguistic variation.

From a synchronic perspective, Medieval French is characterised by a great variability. First of all, it has a free word-order and null subjects. Latin had a nominal declension to help determine the syntactic function of words, but it started to decline very early in Medieval French, and soon became inefficient, as well as rich verbal endings. Buridant (2010) gives this example from *Lancelot, the Knight of the Cart*:

(1) Lancelot Vit la dame de la maison
Lancelot saw the lady of the house

where both *Lancelot* and *la dame de la maison* are candidates for the subject position. The context helps determine their roles: it is the lady who is subject of the sentence. But the grammar we are developing does not have the capability of deducing dependencies from the context.

Furthermore, many dialects are included, which have an impact on the frequencies of occurrences of word forms and syntactic phenomena. The spelling of words was not fixed, even in a same dialect, which leads to the presence of different writings of the same words in a single text. This makes their recognition and analysis more complex. Finally, due to the different forms and domains of texts and the individual styles of authors, different frequencies of phenomena and words can be observed. This should also be taken into account while choosing the datasets to train a disambiguation model. Unlike contemporary languages, synchronic variation in historical texts can be difficult to define because there is no "standard language" we can describe first and extend with the specificities which are encountered in other texts. The number of resources is also limited, which may cause biases in an analysis.

From a diachronic perspective, texts are also subjected to variation. Frequencies of the different word-orders and constructions have evolved through time. These evolutions are however not linear. For example, the OSV order was very rare in the 13th century, and it peaked in the 14th and 15th centuries (Marchello-Nizia, 2008; Combettes and Prévost, 2015).

The valency of some words (i.e. the number and types of argument they require), especially verbs, has evolved too. For example, *morir* (to die) could be a transitive, meaning in this case "to kill", but it is strictly intransitive in contemporary French. Evolution of word sense and use can be observed within Medieval French and has an impact on syntactic analysis because their distributions are different at each period.

Variation is a salient property of Medieval French. It appears at many levels and needs to be handled by parsers.

[3] Some characteristics, like dialect, are more discriminative than others.

4 Related work

Several treebanks for ancient languages are available, for example Latin (Bamman and Crane, 2006), Old English (Taylor, 2007), Medieval Portuguese (Rocio et al., 2003) and Middle High German (Hinrichs and Zastrow, 2012). Other annotated corpora can be found in the CLARIN Research Infrastructure (Hinrichs and Krauwer, 2014). They can be diachronic, which makes the annotation challenging because of morphological and syntactic changes.

The MCVF treebank is the biggest treebank for Medieval French, with 361.283 words for Old French alone, against 251.000 in the SRCMF treebank. Although adjustments must be done in order to adapt the annotation scheme to ours, its size and the presence of texts of Middle French make it a promising resource for machine learning techniques, some of which seem more appropriate for this kind of data.

For example, transfer learning is nowadays used for low-resource languages (Agić et al., 2016). Provided that we develop a parallel corpus for Medieval French, this technique can be explored. It is still possible to do cross-language transfer without such parallel data, as Scrivner and Kübler (2012) did for Old Occitan, a language from the South of France and close to Old French. They chose modern Catalan as their source language for syntax because the word-order is "relatively free", as in Old Occitan. We can use a treebank of Contemporary French, but it is likely to introduce a bias in favour of an analysis close to the modern language. We would not be able to constrain the syntactic models according to linguistic knowledge.

We can also consider using automatic normalisation as an additional layer of annotation, because it has shown efficient for historical texts (Bollmann and Søgaard, 2016). This too is to be explored in a machine learning approach.

This work focuses on a symbolic approach, which will be compared to statistical parsing later on. Brants et al. (2002) pointed out an advantage of parsing with a grammar: the annotation is consistent and has high accuracy. Some projects were successful in adapting existing systems to former stages of a language, as discussed earlier. The extension of the LinGO Redwoods treebank should also be mentioned (Oepen et al., 2004; Toutanova et al., 2005). The authors use a HPSG grammar for analysis and statistical models for disambiguation, ensuring the coherence of annotation. Our grammar should also enable us to annotate new texts following the existing treebank's scheme.

5 Solutions for syntactic analysis

In order to parse Medieval French, we chose to adapt FRMG because of the modularity and flexibility a metagrammar provides.

5.1 French Metagrammar

A metagrammar (Candito, 1996) consists of a hierarchy of small classes describing the rules underlying a grammar. It is a mean to factorise linguistic description, therefore making maintenance and corrections easier. A first general description of a phenomenon is written in a "mother class", from which more specific classes inherit. The metagrammar is compiled into a grammar, which is then used by a parser.

FRMG is a metagrammar based on the Tree Adjoining Grammar (TAG) formalism (Joshi et al., 1975), extended with feature structures. These grammars use elementary trees as units, which have a finite depth and are associated with an item of the lexicon. They are combined to build whole sentences using a non-contextual operation, substitution, and a contextual one, adjunction. A TAG is midly context-sensitive. In FRMG's implementation of TAGs (Villemonte de la Clergerie, 2010), some operators have been added, like disjunction, Kleene star (mainly for coordinations), or guards, expressing conditions on nodes. Sibling nodes are not ordered, which is useful for the analysis of a language with a free word-order.

FRMG's feature structures are hypertags (Kinyon, 2000), a unique structure containing the information of the elementary trees a word can anchor. This is equivalent to a set of supertags. They include grammatical category, sub-categorisation and semantic type. The Lefff lexicon has compatible hypertags with FRMG, which enables the metagrammar to request the information needed for the syntactic analysis, such as POS-tag, gender, number, valency, and the possible forms of the expected arguments of

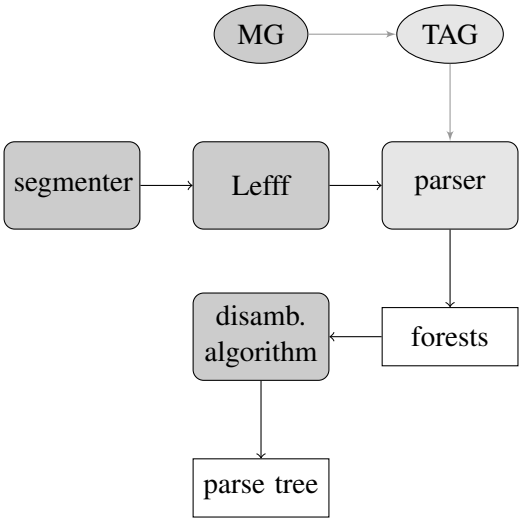

Figure 1: Architecture of the pipeline developed by Boullier et al. (2005)

words (verbs, nouns, adjectives...). If an elementary tree is incompatible with the processed sentence, it is discarded. Having all information available for each word does not cause too much ambiguity.

After the segmentation (see Fig. 1, Boullier et al. (2005)), word lattices are enriched with information from the morphological and syntactic lexicon. This enables to preserve ambiguities. The parser produces all possible analysis in the form of a shared forest of derivation trees, then converted into dependency trees. The disambiguation model selects the most probable solution. It is trained on the French Treebank (FTB) (Abeillé et al., 2003) for contemporary French (Villemonte de la Clergerie, 2013). We will use the SRCMF for our system. Following Guibon et al. (2015), training data should be split according to the metadata of texts, so that the weights in the model fit the new texts to parse. We also consider automatically classifying texts, which would help making use of texts with uncertain metadata or no assigned dialect.

5.2 Adapting FRMG

We use the pipeline described above to parse Medieval French. OFrLex, a lexicon similar to the Lefff (Sagot, 2010), is under development (Sagot, 2019). It includes a new kind of information to add to entries: spelling variants. All variants of a word are linked to it, which is useful for a language with no strict notion of "orthography".

The adaptation of FRMG to Medieval French is a work in progress divided into four main steps. We chose to develop only one metagrammar for the whole period because it is not possible for us to accurately describe each state of language separately. There is no clear boundary between them, they tend to overlap. We consider at first that their main difference is the distribution of frequencies of words and syntactic constructions. As language evolution is not linear, some declining phenomena may rise again some decades later, preventing a straight-forward modelling of language change. Medieval French may be considered as a succession of states of language preparing the contemporary French, but with much more variation and some looser rules, as it can be observed for verbal agreement. Some nouns can either be considered as singular or plural because of their nature as "collective", like *gent* (people).

ex. from *Alexis*: *crient la gent*, transl. "people scream"

(2) crient la gent
VERB (pl) DET (sg) NOUN (sg)

ex. from *Roland*: *La gent de France iert blecee e blesmie*, transl. "The people of France were hurt and turned pale"

(3) La gent de France iert blecee ...
DET (sg) NOUN (sg) ADP NOUN **VERB (sg)** VERB (sg) ...

Our first step towards the adaptation of FRMG is therefore to loosen these constraints, at least for

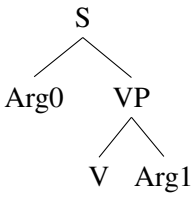

Figure 2: Organisation of the main constituents in FRMG

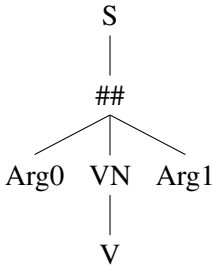

Figure 3: Organisation of the main constituents in our metagrammar. The node ## permits a free order between its children nodes.

collective nouns. Since the states of language are subjected to great variation, all possible analysis should be enabled by the metagrammar until we find out new specific constraints. Missing descriptions also need to be added. We want to be as close to FRMG as possible, keeping most of the descriptions and all the types, to build a continuum between the states of language.

Secondly, in order to deal with free word-order, we chose to change the description of the main constituents. FRMG has a traditional tree representation of a canonical sentence (see Fig. 2), while we chose a flatter representation (see Fig. 3), as advocated by Abeillé et al. (2003). The extension of the TAG formalism permits free order between sibling nodes, which makes our descriptions simpler. Otherwise, we would have to create multiple attachments for verbal arguments in the sentence tree. For example, we find SVO order in *Yvain*, as analysed in the SRCMF:

(4) messire Gauvains ainme Yvain
my lord Gauvain likes Yvain

In the same text, we also find VSO order, which is analysed with the same elementary tree (see Fig. 4):

(5) ainme ele li
likes she him

Thirdly, we want to develop a new mechanism to handle language variation. After all possibilities of analysis are described, we want to restrain some constructions according to the metadata of texts, like the dialect, the genre or the period. The date of a text is particularly informative about the syntax. Some

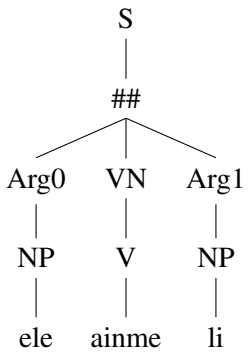

Figure 4: Analysis of sentence (5)

syntactic constructions are known to appear or disappear at a certain period. For example, the OSV order was possible only with a subject pronoun until the 13th century (Schøsler, 1984). By including such a constraint in the metagrammar, we reduce the ambiguity on many sentences. Specificities of dialects have also been described in previous work. For instance, object clitics are usually found before the verb. Some are however found after, but only in texts written in *picard*, a dialect from the North, as in this example from *Escouffe*, v. 4954-55, as cited by Buridant (2010):

(6) prestés me huimais L'ostel
offer me for today hospitality

These special rules can be found in traditional grammars, but we plan to search for new ones with error mining (Sagot and Villemonte de la Clergerie, 2006). Adding a facet handling these exceptions will enable us to describe a general case, instead of under-specifying descriptions in order to enable all possible realisations.

6 Perspectives

We want to extend the SRCMF with the highest quality. Annotation should remain coherent with its annotation scheme. For this purpose, we are currently adapting a large coverage grammar to Medieval French. It has to be completed to be evaluated on a whole corpus and not only on single sentences. This system will then be compared to statistic and neural approaches. This work also aims at developing a methodology for the analysis of heterogeneous data in general, such as tweets and forums.

Acknowledgements

This work takes place in the ANR-16-CE38-0010 PROFITEROLE project (2017–2020), directed by Sophie Prévost.

References

Anne Abeillé, Lionel Clément, François Toussenel. 2003. Building a treebank for French. In *Treebanks: Building and Using Parsed Corpora*. Kluwer Academic Publishers.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter and Anders Søgaard. 2016. Multilingual Projection for Parsing Truly Low-Resource Languages. *Transactions of the Association for Computational Linguistics*.

David Bamman and Gregory Crane. 2006. The design and use of a Latin dependency treebank. *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, p. 67–78.

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

Marcel Bollmann and Anders Søgaard. 2016. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In *COLING*.

Pierre Boullier, Lionel Clément, Benoît Sagot and Eric Villemonte de la Clergerie. 2005. Chaînes de traitement syntaxique. In *TALN 05*, p. 103–112. Dourdan, France.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. *Proceedings of the workshop on treebanks and linguistic theories*, vol. 168.

Claude Buridant. 2010. *Nouvelle Grammaire de l'ancien français*. Paris: Sedes.

Marie-Hélène Candito. 1996. A principle-based hierarchical representation of LTAGs. *Proceedings of COLING-96*. Copenhagen, Denmark.

Bernard Combettes and Sophie Prévost. 2015. La disparition du schéma V2 en français : le rôle de l'opposition marqué / non marqué dans le domaine syntaxique. In *Disparitions. Contributions à l'étude du changement linguistique*, T. Verjans and C. Badiou-Monferran. p. 283–301. Paris: Champion.

Gaël Guibon, Isabelle Tellier, Mathieu Constant, Sophie Prévost and Kim Gerdes. 2014. Parsing Poorly Standardized Language Dependency on Old French. *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, V. Henrich and E. Hinrichs and D.de Kok and P. Osenova and A. Przepiórkowski, p. 51–61. Tübingen, Germany.

Gaël Guibon, Isabelle Tellier, Sophie Prévost, Mathieu Constant and Kim Gerdes. 2015. Searching for Discriminative Metadata of Heterogenous Corpora. *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*.

Erhard Hinrichs and Thomas Zastrow. 2012. Automatic Annotation and Manual Evaluation of the Diachronic German Corpus TüBa-D/DC. *LREC*, p.1622–27.

Erhard Hinrichs and Steven Krauwer. 2014. The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, p.1525–31.

Aravind K. Joshi, Leon S. Levy and Masako Takahashi. 1975. Tree adjunct grammars. *Journal of Computer and System Sciences*.

Alexandra Kinyon. 2000. HYPERTAGS: Beyond POS Tagging. *Natural Language Processing NLP 2000*, D. N. Christodoulakis. Springer-Verlag Berlin Heidelberg.

Christiane Marchello-Nizia. 2008. *L'évolution de l'ordre des mots en français : Chronologie, périodisation, et réorganisation du système*, in Congrès Mondial de Linguistique Française, Paris.

France Martineau. 2008. Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus*, 7, `http://journals.openedition.org/corpus/1508`.

Stephan Oepen, Dan Flickinger, Kristina Toutanova and Christopher D. Manning. 2004. LinGO Redwoods: A Rich and Dynamic Treebank for HPSG. *Research on Language and Computation*, vol. 2, p. 575–596.

Sophie Prévost and Achim Stein. 2013. *Syntactic Reference Corpus of Medieval French (SRCMF)*, version 0.92. `http://srcmf.org`. ENS de Lyon; Lattice, Paris; ILR University of Stuttgart.

Vitor Rocio, Mário Amado Alves, J. Gabriel Lopes, Maria Francisca Xavier and Graça Vicente. 2003. Automated Creation of a Medieval Portuguese Partial Treebank. *Treebanks: Building and Using Parsed Corpora*, Anne Abeillé. Kluwer Academic Publishers.

Benoît Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *7th international conference on Language Resources and Evaluation (LREC 2010)*.

Benoît Sagot. 2019. Développement d'un lexique morphologique et syntaxique de l'ancien français. *26ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*. Toulouse, France.

Benoît Sagot and Eric Villemonte de la Clergerie. 2006. Error mining in parsing results. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*.

Lene Schøsler. 1984. *La déclinaison bicasuelle de lancien français : son rôle dans la syntaxe de la phrase, les causes de sa disparition*. Odense University Press.

Olga Scrivner and Sandra Kübler. 2012. Building an old Occitan corpus via cross-Language transfer. *KONVENS*, p. 392–400.

Ann Taylor. 2007. The YorkTorontoHelsinki Parsed Corpus of Old English Prose. *Creating and Digitizing Language Corpora: Volume 2: Diachronic Databases*.

Kristina Toutanova, Christopher D. Manning, Dan Flickinger and Stephan Oepen. 2005. Stochastic HPSG Parse Disambiguation using the Redwoods Corpus. *Research on Language and Computation*, vol. 3, p. 83–105.

Eric Villemonte de la Clergerie. 2005. From metagrammars to factorized TAG/TIG parsers. *Proceedings of the Ninth International Workshop on Parsing Technology - Parsing '05*.

Eric Villemonte de la Clergerie. 2010. Building factorized TAGs with meta-grammars. *The 10th International Conference on Tree Adjoining Grammars and Related Formalisms - TAG+10*.

Eric Villemonte de la Clergerie. 2013. Improving a symbolic parser through partially supervised learning. *The 13th International Conference on Parsing Technologies (IWPT)*. Naria, Japan.