Parsing Poorly Standardized Language Dependency on Old French

Gaël Guibon, Isabelle Tellier, Mathieu Constant, Sophie Prévost, Kim Gerdes

Parsing Poorly Standardized Language Dependency on Old French

Gael Guibon (1), Isabelle Tellier (1,2)
Matthieu Constant (3), Sophie Prevost (1) and Kim Gerdes (2,4)

(1) Lattice CNRS
(2) université Paris 3 - Sorbonne Nouvelle
(3) université Paris-Est, LIGM
(4) LPP CNRS
E-mails: gael.guibon@gmail.com, isabelle.tellier@univ-paris3.fr,
Matthieu.Constant@u-pem.fr, sophie.prevost@ens.fr, kim@gerdes.fr

Abstract

This paper presents results of dependency parsing of Old French, a language which is poorly standardized at the lexical level, and which displays a relatively free word order. The work is carried out on five distinct sample texts extracted from the dependency treebank *Syntactic Reference Corpus of Medieval French* (SRCMF). Following Achim Stein's previous work, we have trained the *Mate* parser on each sub-corpus and cross-validated the results. We show that the parsing efficiency is diminished by the greater lexical variation of Old French compared to parse results on modern French. In order to improve the result of the POS tagging step in the parsing process, we applied a pre-treatment to the data, comparing two distinct strategies: one using a slightly post-treated version of the TreeTagger trained on Old French by Stein, and a CRF trained on the texts, enriched with external resources. The CRF version outperforms every other approach.

1 Introduction

Today's research on historic language data is still profoundly different from usage based analyses of modern languages. Historic language data are generally sparse and intrinsically inhomogeneous. Common statistical corpus analysis methods are thus poorly suited and less successful as even simple frequency counts on raw corpora fail to provide reliable results. Moreover, one central goal of diachronic linguistics is the analysis of structural change over time, which is a gradual process calling for quantitative methods. However, very few resources of historic language are available in digital formats, and even fewer are provided with any type of annotation that could allow the application of standard corpus linguistic methods. There

is a variety of reasons for this situation, ranging from epistemological difficulties to the lack of economic interest. In this paper, we address the technical problems of producing, extending, or consolidating these resources with the help of statistical parsers.

Treebank development is often made easier and more precise by the use of machine learning techniques in a bootstrapping approach. Today's successful machine learning techniques rely on the underlying assumption that word forms are spelled the same way with only few exceptional and thus unknown forms whose analysis can be guessed correctly from the context. In this paper, we explore how difficult dependency parsing on non-standardized text actually is, also compared to equivalent tasks on more homogeneous texts of modern languages.

The treebank we use for these measures is the manually annotated *SRCMF* treebank (Syntactic Reference Corpus of Medieval French)[1] [10]. We explore at which point in the standard incremental parsing setup (lemmatization, POS-tagging, dependency parsing) the inhomogeneous character of the data interferes most strongly. In particular, two strategies are tested for POS-tagging to overcome this difficulty: one based on a slightly post-treated version of the TreeTagger trained on a large corpus of Old French, the other applying Conditional Random Fields (CRF) learning for various distinct texts separately. We show that CRFs allow to greatly improve previous results.

In the following, we first introduce the SRCMF treebank (section 2), and the portions of it we have used. We also provide some indicators to quantify its variability relatively to contemporary French. We then briefly present a related work (section 3). We finally explain the experiments conducted to minimize the impact of the lack of standardization on the final parsing quality (section 4).

2 Presentation of the Corpus

2.1 General presentation

Our research is based on the *Syntactic Reference Corpus of Medieval French* (SRCMF) [10], a heterogeneous treebank of Old French which was developed in a joint research project funded by the *Deutsche Forschungsgemeinschaft*[2] and the *Agence Nationale de la Recherche*[3] (ANR) from March 2009 to February 2012. The origin of this project was a collection of important medieval French texts whose electronic versions are stemming from the *Base de Francais Médiéval*[4] (BFM) [4] and the *Nouveau Corpus d'Amsterdam*[5] (NCA) [5]. It has been built to serve as a reference treebank of Old French.

[1] http://srcmf.org/

[2] http://www.dfg.de/

[3] http://www.agence-nationale-recherche.fr/

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

[5] http://www.uni-stuttgart.de/lingrom/stein/corpus/

Text	Date	Nb words	Nb sent.	Type
Chanson de ***Roland***	1100	29 338	3843	verse
Yvain by Chretien de Troyes	1177-1181	42 103	3735	verse
La ***Conqueste*** *de Constantinople* by Robert de Clari	>1205	33 994	2282	prose
Queste del Saint ***Graal***	1220	40 000	3049	prose
Aucassin *et Nicolete*	late 12c.-early 13c.	9387	985	verse & prose

Table 1: Texts from the SRCMF used in our experiments

Although the original texts contained few punctuations or other indications of segmentation, they have been segmented into clauses made around a finite verb. These clauses will be referred to as "sentences" in the following, even if a subordinate clause is not exactly a sentence. The original electronic versions of the texts already came with a POS tagging (50 POS tags), whereas the fine-grained dependency annotation (31 syntactic functions) was added manually, using Mazziotta's tool *NotaBene* [8].

In SRCMF, the POS tags were verified and each clause was syntactically analyzed by means of a dependency tree. Only *Yvain* includes a manually verified lemmatization.

From the SRCMF we choose five texts, shown in Table 1, of different periods, genres, and dialects. The first four of these texts are similar in size and date from the early 12th to the 13th century, written either in prose or verse. By way of comparison, the fifth selected text has a different size and is composed of a mix of verse and prose. The texts differ in the regional dialect they have been written in: Norman for *Roland*, Champenois for *Yvain*, Picard for *La Conqueste* and *Aucassin*, while *Graal* is unmarked for regional dialect. In fact, our experiments also include other texts, but the results for these five texts are representative of the whole results.

2.2 Heterogeneity of the corpus

The main reason for the heterogeneous character of the data is not so much the time span in which the different texts have been produced (120 years), but rather the lack of spelling norms, which only gradually developed historically under the influence of printing and the emergence of grammar books. In the middle ages, each author could develop their own written preferences influenced by their dialect. However, medieval texts display important variations which correspond not only to the dialects, since some spelling variants between texts belonging to the same dialect can also be observed. Still more surprisingly, even a single text by the same author may display spelling variations in some words. Table 2 provides some examples of words appearing in various forms in the same text (*Yvain*), whereas only a single form persists in contemporary French.

In order to measure SRCMF's word form variability, we can compare it with

Contemporary word	Form variations
Bretagne	bretaingne | bretaigne
vilain(e)(s)	vilains | vileins | vilainne
ainsi	ensi | einsi | ainsi

Table 2: Examples of word form variability

contemporary French. Unfortunately, we only have verified lemmas for one corpus (*Yvain*, from Chretien de Troyes), so we can only study the word form variability of a small part of SRCMF and we cannot quantify the differences due to the various kinds of dialects, authors, or even centuries. We used the French Treebank (FTB) [6] [1] as a sample of contemporary French. For both corpora, we computed the number of distinct forms corresponding to a single lemma, and averaged this number for each distinct POS tag. Table 3 shows the values obtained for the main morpho-syntactic categories. This indicator allows us to quantify the variability of spelling, at least for *Yvain*.

POS \ Corpus	French Treebank	SRCMF's Yvain
proper noun	1	1.25
common noun	1.31	1.31
infinitive verb	1	1.10
finite verb	2.48	3.15
determinant	1.06	2.21
adjective	1.63	1.68
adverb	1.01	1.40
Average number of forms per lemmas	1.57	2.25

Table 3: average number of forms for a lemma, for the FTB and *Yvain*

As expected, the values for *Yvain* are always higher than those for the FTB. Some categories of words which are nowadays considered as invariable (proper nouns, infinitive verbs) can correspond to various forms in *Yvain*. For example, the name *Yvain* itself can appear under four different forms: *Yvains*, *Yveins*, *Yvain*, and *Yvein*. This name being the main character of the text, it shows how poorly standardized Old French can be.

3 Previous works

Training statistical parsers is becoming a common step in linguistic ressource development in general and treebank construction in particular, often mixed with manual and rule-based approaches. But we believe that it is an interesting endeavor in itself, widely under-used, as a tool of linguistic analysis because it can

[6] http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php

provide information about the consistency of the syntactic annotation or about the variability of different sub-corpora. Cross-training a parser on a sub-corpus and applying the resulting parser on another corpus gives interesting insights not only in the difficulties of parsing these heterogeneous texts, but also in the historic and genre differences between these texts, as well as the dialectal spelling differences.

Achim Stein first conducted such syntactic parsing experiments on the SRCMF corpus [9]. The results are (partially) reported in Table 4. For this work, he trained the TreeTagger[7] on the Old French *Nouveau Corpus d'Amsterdam* in order to obtain POS and lemmas, and used Bernd Bohnet's dependency parser *Mate parser* [2] for the syntactic analyses. Although only about 70% of the graphemic forms received lemmas, many of which are ambiguous, this unverified lemmatization was used as the initial data for the parsers. The results of cross-training a parser are reported below.

Train \ Test		Auc.	Rol.	Graal	Yvain	Conq.
Aucassin	UAS		63.84	70.23	63.57	74.00
	LAS		44.56	57.16	48.04	61.88
Roland	UAS	67.73		71.03	64.48	67.80
	LAS	52.93		57.67	49.71	55.07
Graal	UAS	75.92	66.87		72.79	76.20
	LAS	63.06	46.67		58.24	64.49
Yvain	UAS	74.71	68.00	80.80		72.27
	LAS	61.96	48.45	70.06		58.68
Conq.	UAS	70.27	61.93	70.53	61.58	
	LAS	56.32	42.00	57.98	45.44	

Table 4: Stein's scorings

Except for this work, the SRCMF has mainly been used for linguistic purposes. We do not refer to these other studies here, as our work is clearly a continuation of Stein's experiments, which serve as our baseline.

4 Our experiments

Our purpose is to improve Stein's results. We expect to obtain a better performance of the Mate parsers by improving the initial POS labeling and lemmatization phases. Thus, we first explain the strategy used to obtain a good POS tagger, then we detail the results obtained for the parsing phase.

4.1 POS Tagging and Lemmatization

In order to achieve a comparable experimental setup with reliable performance measures, we produced sample extracts similar in size as the ones used in the pre-

[7] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

vious experiments: approximately 16000 words per sample. *Aucassin* contains only 9387 words but it was kept since it was the only text containing both verse and prose, which allowed us to see if this implied different results. Just as Stein, we used the *Mate* dependency parser, but we used it only for the dependency annotation. For the preliminary POS tagging, we tried two different strategies:

- Tagging the data with the *TreeTagger* trained by Stein. We also applied basic surface rules that remove useless tagger outputs and improve the lemmatization. When a lemma is not recognized, the word is associated with a specific "<nolem>" string.
- Training a specific POS tagger with Conditional Random Fields (CRF) [6], implemented by *Wapiti* [8] [7] for each training text separately. CRFs allow to take into account various types of contextual and external information such as the (slightly post-treated) *TreeTagger* lemmatization results and a lexicon that we extracted from the BFM corpus. This lexicon associates to each word present in the BFM corpus the set of its possible POS tags. The feature templates defined in this CRF using these external resources take the following forms:
  - check whether the current word (resp. the previous word, resp. the next word) is associated with the <nolem> lemma value by the TreeTagger
  - check the value of the lemma predicted by TreeTagger for the current word (resp. the previous word, resp. the next word)
  - check the value of two consecutive lemmas predicted by TreeTagger (for the previous and the current words, for the current and the next words)
  - for each distinct POS tag, check whether it can be associated with the current word in the BFM lexicon
  - concatenate all the distinct POS tags associated with the current word in the BFM lexicon

  Other features were used, such as checking the near contextual words, the final letters of the words (up to 4 letters), the lowercase value of the words, whether or not word forms begin by an uppercase, whether word forms begin neither by a letter nor by a number, whether the word's final letter is a special character (e.g an apostrophe for elision).

The main advantage of CRFs is to take into account more external resources and more contextual information, which appears to be crucial for a POS labeling of good quality for our historic language data. While *TreeTagger* uses a model trained on the *Nouveau Corpus d'Amsterdam*, with CRFs (which require fewer training data) we treat each text separately. Tables 5 and 6 display the accuracies

[8] `http://wapiti.limsi.fr/`, we used the 1.4.0 version

of the various POS taggers obtained on our data. CRFs usually obtain far better results than the TreeTagger.

Train \ Test	Auc.	Rol.	Graal	Yvain	Conq.
Aucassin		80.00	85.76	80.03	87.86
Roland	80.48		82.66	78.20	84.13
Graal	85.38	80.58		82.70	86.84
Yvain	83.13	80.22	89.05		82.11
Conq.	80.48	74.51	79.98	71.04	

Table 5: Accuracies of cross-trained POS taggers learned by CRF

Test	Accuracy
Aucassin	70.94
Roland	71.59
Graal	84.28
Yvain	66.76
Conq.	65.65

Table 6: Accuracies of the POS produced by the TreeTagger

4.2 Parsing

As previously mentioned, like Stein, we used the *Mate parser* (anna-3.61 version) for the syntactic analysis of our texts. For each experiment, we tried two distinct POS labeling strategies: either the (slightly post-treated) *TreeTagger* trained by Stein, or a specific POS tagger learned on the training text by a CRF. In each case, the lemmatization was provided by the TreeTagger, improved by surface rules. The "cross learning" results we obtained are shown in Table 7 and Table 8.

Train \ Test		Auc.	Rol.	Graal	Yvain	Conq.
Aucassin	UAS		66.12	73.57	68.67	76.02
	LAS		49.34	60.96	51.10	64.84
Roland	UAS	71.30		72.00	68.08	69.20
	LAS	58.36		62.61	54.16	54.80
Graal	UAS	75.34	67.40		72.84	77.21
	LAS	66.38	51.27		61.11	66.01
Yvain	UAS	74.67	69.46	81.05		73.83
	LAS	64.06	50.16	70.51		61.32
Conq.	UAS	72.07	65.20	71.08	62.37	
	LAS	59.65	45.33	60.18	48.04	

Table 7: Syntactic analysis results with the POS produced by the TreeTagger

Train \ Test		Auc.	Rol.	Graal	Yvain	Conq.
Aucassin	UAS		76.27	79.00	72.70	79.20
	LAS		58.98	65.02	57.63	68.50
Roland	UAS	72.26		73.02	70.64	73.86
	LAS	56.84		58.45	55.19	61.27
Graal	UAS	78.48	77.82		75.16	80.88
	LAS	65.52	59.79		61.15	69.08
Yvain	UAS	77.07	79.20	82.42		76.74
	LAS	64.72	61.58	70.41		63.81
Conq.	UAS	75.02	72.85	76.03	66.07	
	LAS	60.59	54.71	61.87	50.14	

Table 8: Syntactic analysis results with the POS produced by the CRFs

As can be seen in these results, the syntactic analyses based on CRF-induced POS tags outperform every other approach, improving Stein's results by nearly 10% in average.

We can see that there is a huge gap between UAS and LAS in our results, as it was the case in Stein's experiments. We suspect that this gap, which is not common in dependency parsing, is due to the size of the dependency label set (around 30 in our case, as compared with around 10 in standard treebanks) and/or the higher rate of variability in Old French, i.e. the fact that there exist several different forms for a same lemma (cf. Table 3).

4.3 Influence of the Lemmas on the Parsing

To evaluate the influence of the lemmas on the parser, we have conducted other experiments using *Yvain*, for which corrected lemmas are available. We have divided our gold version of *Yvain* with verified lemmas into two different sub-corpora of about 16 000 words each, one dedicated for training the parser, the other for testing it. We did the exact same division on *Yvain* with *TreeTagger* predicted lemmas.

	TT predicted lemmas	Verified lemmas
lemma acc.	58.84	100
UAS	88.99	89.36
LAS	79.55	80.53

Table 9: Lemmas'influence on the parsing (with gold POS)

The experiment whose results are displayed in table 9 only shows the lemmas' influence on dependency parsing, not on POS tagging. Here, as opposed to our previous works, the training and testing corpora are extracted from the same text (the only one provided with verified lemmas) and the parser could take advantage of gold POS tags, which explains why the parse scores are much higher than in

previous experiments, even with a small training set. The 1% improvement in LAS of this experiment confirms that lemmas have an influence on the parser quality, even without considering their influence on correct POS tags (which indirectly appear in the final parse results).

We can also compare our results to a simple baseline of dependency parsing using *Mate* for various portions of the French Treebank (FTB) with gold POS (some of them similar in size to our training set from *Yvain*). We obtain, without any further treatments, the scores in Table 10.

In fact, these results are hard to compare, for the following reasons:

- in the variant of the FTB we used, multi word units are not pre-treated: they have to be recognized during the parsing phase, which is a harder task than just parsing. In fact, the current state of the art for the dependency parsing on contemporary French with pre-recognized multi word units can reach 90.3% in UAS and 87.6% in LAS [3].
- the average length of a "sentence" in *Yvain* is about 10 words, while it is about 30 for the FTB, which implies that the task of parsing the FTB is much more difficult, as the syntax of the sentences it contains is more complex

It is nevertheless possible to draw several conclusions from these experiments. First, as already known, the training corpus size is very important to obtain high scores in parsing. But we could not obtain a large size corpus with perfect lemmas for Old French. Secondly, we can see that, for the available quantity of training data, the results obtained by the trained parser are already not bad. This means that, at the syntactic level, Old French is "regular" enough for training a parser.

	FTB train $\approx$ 450000 w.	FTB train $\approx$ 16000 w. (average of 5 experiments)
UAS	88.80	81.19
LAS	85.58	76.04

Table 10: Baseline on the French Treebank using *Mate*

5 Conclusion

In this paper, we have explored the difficulties and possible improvements of statistical parsing of a poorly standardized language. The results of our experiments show that a main issue in the process is the lack of a correct lemmatization, which percolates through the whole parsing process and is partly responsible for the final parsing quality. We managed to get around the poor lemmatization by trying to directly influence the quality of POS tagging and by doing so we obtained far better results than what has been achieved previsously on Old French. Adding external resources seems to be one of the keys to increase the final score. With these experiments, we managed to obtain a correct parsing quality, which could provide a

reasonable base for manual correction, but the results remain well below the level of the scores obtained for contemporary more standardized languages. Note, however, that our training sets of dependency trees were relatively small compared to other treebanks.

In order to obtain a better comparison we decided to test on corpora of approximatively the same size. These experiments confirmed again that a poorly standardized corpus results in a huge drop on LAS scoring. Moreover, the fact that the results show similar scores in UAS can be analyzed as a symptom of the higher variability in Old french.

In the present state of our experiments, it remains difficult to draw some solid linguistic conclusions. The relatively good scores when training *Aucassin* on *Roland* is somewhat unexpected, since *Aucassin* is far later than *Roland* and partially written in prose (whereas *Roland* is written in verse). Both also differ in genres. *Conqueste* is known to be somewhat untypical with regard to some syntactic features, as well as rather marked from a lexical and morphological point of view, which could explain the fact that we obtain worse scores with it than with the other texts. *Graal* and *Yvain*, though differing in their form, are not very distant in time, and moreover they share some common literary themes: this could explain the relatively good scores.

These brief linguistic conclusions certainly deserve further investigation. The asymmetries of our cross-trained tables should be further analysed. It is also still not clear whether the efficiency of a parser trained on one text and applied to another one is correlated with the historic proximity of the writing period, with the texts' genres, or more basically simply with the texts' length (remember that *Aucassin* is smaller than the other texts). If it appears to be linguistically relevant, the results of cross-training a syntactic parser could be used as a distance measure between genres and origin time of texts.

Note also that the variability explored here is mainly of a lexical nature. Only a serious study of the syntactic variations (e.g. word order of Old French is freer than in contemporary French) and its influence on the machine learning process could improve the scope of the results.

References

[1] Anne Abeillé, Lionel Clément, and François Toussenel. Building a treebank for french. In *Treebanks*, pages 165–187. Springer, 2003.

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

[3] Marie Candito, Benoît Crabbé, Pascal Denis, et al. Statistical french dependency parsing: treebank conversion and first results. In *Proceedings of*

the Seventh International Conference on Language Resources and Evaluation (LREC 2010), pages 1840–1847, 2010.

[4] Céline Guillot, Alexei Lavrentiev, and Christiane Marchello-Nizia. Document 2. les corpus de français médiéval : état des lieux et perspectives. *Revue française de linguistique appliquée*, XII:125–128, 2007.

[5] Pierre Kunstmann and Achim Stein. Le nouveau corpus d'amsterdam. In *"Actes de l'atelier de Lauterbad"*, pages 9–27, 2007.

[6] John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pages 282–289, Seattle, Washington, 2001.

[7] Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics, July 2010.

[8] Nicolas Mazziotta. Logiciel notabene pour l'annotation linguistique. annotations et conceptualisations multiples. In *Recherches qualitatives. Hors-série "Les actes"*, volume 9, 2010.

[9] Achim Stein. Parsing heterogeneous corpora with a rich dependency grammar. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[10] Achim Stein and Sophie Prévost. Syntactic annotation of medieval texts: the syntactic reference corpus of medieval french (srcmf). In Tübingen: Narr, editor, *New Methods in Historical Corpus Linguistics*. 2013. Corpus Linguistics and International Perspectives on Language, CLIP Vol. 3, P. Bennett, M. Durrell, S. Scheible and R. Whitt (eds).