Searching for Discriminative Metadata of Heterogenous Corpora

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

Searching for Discriminative Metadata of Heterogenous Corpora

Gael Guibon (1), Isabelle Tellier (1,2)
Sophie Prevost (1), Matthieu Constant (3) 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

November 25, 2015

Abstract

In this paper, we use machine learning techniques for part-of-speech tagging and parsing to explore the specificities of a highly heterogeneous corpus. The corpus used is a treebank of Old French made of texts which differ with respect to several types of metadata: production date, form (verse/prose), domain, and dialect. We conduct experiments in order to determine which of these metadata are the most discriminative and to induce a general methodology.

1 Introduction

Labeled data used to train POS taggers or syntactic parsers by supervised machine learning technics are usually rather homogenous: The texts they come from share a common origin and most of their metadata. Yet, most actual text collections that NLP tools have to handle today are heterogenous in many ways. The consequence of this inadequacy is well known: Programs trained on homogenous texts by supervised machine learning approaches do not perform well when applied to new texts that differ from them in any important aspects, such as subject, genre or sublanguage. To address this problem, domain adaptation has become an important issue in machine learning NLP.

In this paper, we explore a very heterogenous corpus of Old French but the problem we tackle is not exactly domain adaptation. We want to use machine learning for *corpus exploration*, i.e. as a way to search for the most discriminative metadata of our texts. As a matter of fact, they belong to a *highly heterogenous*

treebank and vary in dialect, domain, production date, and form (verse and prose). In this context, new questions arise: Which of these properties (metadata, in the following) influence most the language in which they are written? How can we best train a POS tagger and a parser on this treebank, in order to annotate a new yet unlabeled text of Old French? Is it better to use, as training data, a small homogenous corpus similar to the new text or a large heterogeneous dissimilar one? These questions can also be relevant for other heterogenous corpora and cross domain applications, so our purpose will also be to provide a general methodology.

2 Syntactic Reference Corpus of Medieval French

The SRCMF [1][11] is a treebank of Old French texts enriched with POS tags (chosen among 60 distinct possible ones) and fine-grained dependency structures (labeled with 31 distinct syntactic functions) which were built manually during an ANR-DFG joint research project (2009-2012). The corpus consists of 15 texts (245 000 words) whose electronic versions are stemming from the "Base de Francais Médiéval" (BFM)[2] [4] and the "Nouveau Corpus d'Amsterdam" (NCA)[3] [5]. The selection of the included texts is based on criteria such as date, form (verse/prose), domain (historical, literary...) and dialect. From the SRCMF we choose 10 texts, whose metadata are shown in Table 1.

Text	Date	Words	Form	Dialect	Domain
Vie Saint ***Légier***	late 10c.	1388	verse	n/a	religious
Vie de Saint ***Alexis***	1050	4804	verse	normand	religious
Chanson de ***Roland***	1100	28 766	verse	normand	literary
Lapidaire *en prose*	Mid. 12c.	4708	prose	anglo-norm.	didactical
Yvain, Chr. de Troyes	1177-1181	41 305	verse	champenois	literary
La ***Conqueste*** *de Constantinople*, R. de Clari	>1205	33 534	prose	picard	historical
Queste del Saint ***Graal***	1220	40 417	prose	n/a	literary
Aucassin *et Nicolete*	late 12c.-early 13c.	9844	verse & prose	picard	literary
Miracles from Gautier de **Coinci**	1218-1227	17 360	verse	picard	religious
Roman de la ***Rose*** from Jean de Meun	1269-1278	19 339	verse	n/a	didactical

Table 1: Texts of the Corpus and their Metadata

The syntactic annotation is based on a dependency model [12, 9], which distinguishes between, on the one hand, syntactic units and, on the other hand, different functions (such as "subject", "object", "adverbial", "auxiliary", "modifier"...),

[1] http://srcmf.org/

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

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

which specify the relation between a head and the structures depending on it.

3 Most Discriminative Metadata in SRCMF

In this section we propose a new experimental strategy to explore SRCMF based on metadata-related experiments, in order to find the metadata that are the most discriminative to tag and parse a new text.

3.1 Protocol of our Experiments

Our experiments are not designed to search for the best parser and configurations, this has already been studied before [8, 3]. But, to evaluate the impact of each metadata, it is important to obtain comparable results. So, we define a general protocol for all the experiments whose results are reported in this paper. For each metadata value, we process as follows:

- The data are split into a training set (of equal size for each possible value) and a test set made of the remaining content;
- On all training and test sets, the lemmas are predicted by *TreeTagger* trained on the *Nouveau Corpus d'Amsterdam* (NCA) parameters[4] [10];
- A CRF part-of speech tagger [6] is trained on the training set using *Wapiti* 1.4.0 [7] with the same templates taking into account the contexts of words and lemmas, word endings, etc. as in [3];
- A Dependency parser is trained on the training set (with Gold POS labels) with *Mate-tools* (anna-3.61 [5]) [1]
- The lemmatizer, the tagger and the parser are successively applied to the test set. The tagger is evaluated by its accuracy, the parser by the classical UAS and LAS values.

3.2 Summary of the Experiments

Dates. We first conduct these experiments using the production dates of the texts, which are usually considered as discriminative metadata. This approach seems all the more obvious for the SRCMF whose texts date from the late 11th to the 13th century, which allows us to make a comparison between centuries. Nonetheless, as the 11th century sub-corpus is too small (less than 7 000 words), we restrict ourselves to the other two. For this time slicing, we also have to remove *Aucassin*

[4] http://bfm.ens-lyon.fr/article.php3?id_article=324

[5] Available at https://code.google.com/p/mate-tools/downloads/detail?name=anna-3.61.jar&can=2&q=

from the data, as it is not clear whether it has been written in the 12th or the 13th century.

The results show that applying a model trained on the same century gives higher scores, especially in LAS. In fact, even if high UAS, LAS, and accuracy match with the correlation of century between train and test sets, the similarity between sub-corpora shows the contrary, with far more shared words when centuries differ, which is quite unexpected. This induces that the Out-Of-Vocabulary rate is not enough to predict the results. To confirm this intuition, we conducted few additionnal iterations of this experiment following the same method, in which texts from a century have been mixed, and a training set of 50 000 words[6] taken at random among them has been built. The performances reached by the models learned from these training sets lead to the same conclusions as previously. It is important to point out that even if both century based sub-corpora are composed of more than one domain and both verse and prose texts, the data from the 13th century are all written in Picard or in an undefined dialect. So, as one century does not vary in dialects, the date value is not independent from other metadata values. Moreover, as stated before, we could only use two centuries, which limits the scope of the results.

	Words	**Units**	**Sentence**	**Sentence length**
12th century [train]	50002	7337	5430	9
13th century [train]	50009	6934	3767	13
12th century [test]	24777	4844	2685	9
13th century [test]	60638	7597	4538	13

Table 2: Characteristics of the Time Sliced Sub-Corpora

Forms. The impact of the production date has been clearly shown but other experiments have to be conducted to compare it with the effect of the other metadata. We now consider text forms, i.e. if they are written in verse or prose. In the SRCMF, most of the texts are in verse but some of them are in prose and one text (*Aucassin et Nicolete*) is written in both forms. It is well aknowledged that verses are syntactically more constrained than prose, and moreover they have a richer lexicon (based on a higher amount of different units): We thus wonder if this could have an impact, in particular on parsing. This effect combines with the fact that sentences in prose are usually longer than those in verse. The results in Table 5 contradict the idea that a different text form induces a significantly different model of dependency parsing. Indeed, the prose test corpus obtains higher UAS and LAS even from the verse model. Compared to using centuries, this time shared lexicon and known words show a greater similarity between corpora of the same text form.

According to our results, text form seems to be useful to discriminate texts from a corpus but only when it comes to part-of-speech tagging, as dependency parsing

[6]Because of the necessity to keep complete sentences, the size of randomly built training sets may vary by few words

Train \ Test		**12th century [test]**	**13th century [test]**
12th century	UAS	**88.81**	83.14
	LAS	**79.91**	71.93
	ACC	**94.69**	89.62
	Unknown | known words	**91.39** | 08.61	78.72 | **21.28**
	Different | shared lexicon	**61.20** | 38.80	28.59 | **71.41**
	Unknown | known words UAS	**81.05 | 90.00**	71.60 | 85.13
	Unknown | known words LAS	**66.42 | 81.47**	54.18 | 75.72
	Unknown | known words ACC	**87.29 | 95.39**	78.14 | 92.73
13th century	UAS	82.24	**89.07**
	LAS	69.24	**80.75**
	ACC	88.67	**94.62**
	Unkonwn | known words	73.83 | **26.17**	**92.25** | 07.75
	Different | shared lexicon	33.96 | **66.04**	**50.12** | 49.88
	Unkonwn | known words UAS	**76.94** | 86.61	74.35 | **88.77**
	Unkonwn | known words LAS	56.75 | 75.84	**57.96 | 80.46**
	Unkonwn | known words ACC	80.13 | 91.69	**85.31 | 95.41**

Table 3: Results Using Time Slicing

results seem irrelevant to our purpose. This makes the form metadata useful, but less reliable than the time slicing, when it comes to corpus exploration.

	Words	**Units**	**Sentence**	**Sentence length**
prose [train]	41910	4320	4320	14
verse [train]	41907	6840	6840	7
prose [test]	36749	4370	4370	12
verse [test]	34478	4417	4417	10

Table 4: Characteristics of the Text Form Based Sub-Corpora

Domains. Moving on to domain-related corpora, we now have more than two corpora to compare.

The notion of "domain" usually corresponds to what the texts are about. In the context of SRCMF, it is more related to the *literary genre* of these texts. Both notions do not exactly coincide but, in both cases, texts from the same domain should share some specific content words. As content words are less frequent and more ambiguous than grammatical words, we expect their presence or absence in both the training and test sets to affect the parsing results.

Domain adaptation is a prolific research field in machine learning and previous works have shown that it is possible to obtain better results by focusing on a specific domain rather than using a global approach [2]. This is why we could expect the domain value to have a great impact on the results.

A training set of about 16 000 words is first extracted from each of our four domain-specific sub-corpora, to ensure balanced training data. Test sets are made of the remaining content (Table 6). As expected, when the training and the test sets come from the same domain, the results, given in Table 7, are (in average) better

Train \ Test		**Prose [test]**	**Verse [test]**
Prose	UAS	**85.47**	76.33
	LAS	**74.96**	62.96
	ACC	**91.36**	83.61
	Unknown | known words	16.49 | **83.51**	**21.26** | 78.74
	Different | shared lexicon	57.02 | **42.98**	**77.05** | 22.95
	Unknown | known words UAS	**73.76 | 87.78**	65.87 | 79.15
	Unknown | known words LAS	**55.48 | 78.81**	46.37 | 67.44
	Unknown | known words ACC	**77.33 | 94.14**	76.78 | 85.46
Verse	UAS	**83.12**	82.79
	LAS	**71.52**	71.40
	ACC	90.06	**90.78**
	Unknown | known words	**18.81** | 81.19	14.03 | **85.97**
	Different | shared lexicon	**66,47** | 33,53	42.52 | **57.48**
	Unknown | known words UAS	**73.43 | 85.37**	72.39 | 84.49
	Unknown | known words LAS	55.45 | **75.24**	**55.62** | 73.98
	Unknown | known words ACC	81.02 | **92.15**	**84.13** | 91.86

Table 5: Verse [train] Shows Better Results on Prose [test] than on Verse [test], while Results on Known Words and Shared Lexicon Suggest the Contrary

and higher scores go along with the proportion of shared lexicon. This suggests that the domain is indeed a discriminative metadata. Moreover, results are quite stable except for the historical trainset, which shows an even greater gap when both sets come from the same domain, with an increase of about 30% in LAS. But, it is proper to remind that the historical corpora is made of only one text (*La Conqueste de Constantinople*).

	Words	**Units**	**Sentence**	**Sentence length**
Didactical [train]	16003	3820	1238	12
Historical [train]	16007	2298	1108	14
Literary [train]	16009	3529	1526	10
Religious [train]	16011	3645	1470	10
Didactical [test]	8013	2374	680	11
Historical [test]	17528	2414	1249	14
Literary [test]	104323	10828	10209	10
Religious [test]	7541	2182	708	10

Table 6: Domain-Specific Corpora's Charact.

Dialects. We then finally evaluate dialects as a discriminative metadata. A dialect speaker understands, at least partly, another dialect of the same language, as dialects share large parts of lexicon and grammar. In Table 9, we aim to determine whether or not the same holds for the dialects of our corpus. We use three sub-corpora based on the three distinct dialects, each training set being approximatively made up of 20 000 words.

We observe a huge increase in performance (of about 10 points in LAS, UAS, and accuracy) while applying a model on a same dialect. When the training and

Train \ Test	**Didactical[test]**	**Historical[test]**	**Literary[test]**	**Religious[test]**
Didactical [train]				
UAS	**81.78**	78.88	80.11	70.05
LAS	**71.23**	67.28	66.67	55.04
ACC	**90.75**	87.58	87.08	80.80
Unknown | known words	16.53 | **83.47**	**31.15** | 68.85	26.08 | 73.92	30.58 | 69.42
Different | shared lexicon	50.19 | **49.81**	78.05 | 21.95	**83.85** | 16.15	69.67 | 30.33
Unknown | known w. UAS	**71.68 | 83.78**	69.12 | 83.29	70.13 | 83.63	59.80 | 74.57
Unknown | known w. LAS	**53.93 | 74.66**	52.89 | 73.79	50.69 | 72.29	38.46 | 62.34
Unknown | known w. ACC	80.89 | **92.70**	**81.30** | 90.43	77.53 | 90.44	66.96 | 86.89
Historical [train]				
UAS	67.49	**90.07**	73.03	32.29
LAS	51.12	**82.20**	57.30	45.08
ACC	72.74	**95.66**	76.67	69.93
Unknown | known w.	41.09 | 58.91	08.08 | **91.92**	38.66 | 61.34	**42.57** | 57.43
Different | shared lexicon	81.94 | 18.06	46.67 | **53.33**	**90.46** | 09.54	79.84 | 20.16
Unknown | known w. UAS	58.08 | 74.05	**80.16 | 90.94**	65.06 | 78.05	52.80 | 69.33
Unknown | known w. LAS	38.24 | 60.11	**63.70 | 83.92**	45.20 | 64.93	31.56 | 55.10
Unknown | known w. ACC	62.67 | 79.77	**87.50 | 96.38**	66.95 | 82.80	57.20 | 79.38
Literary [train]				
UAS	77.22	82.02	**84.79**	73.09
LAS	64.07	70.79	**73.63**	59.01
ACC	85.10	88.95	**91.93**	83.25
Unknown | known w.	27.01 | 72.99	**27.35** | 72.65	14.42 | **85.58**	27.16 | 72.84
Different | shared lexicon	68.17 | 31.83	73.58 | 26.42	**75.36** | 24.64	65.96 | **34.04**
Unknown | known w. UAS	66.17 | 81.31	74.07 | 85.02	**74.28 | 86.56**	61.18 | 77.53
Unknown | known w. LAS	46.03 | 70.74	**57.21** | 75.90	56.25 | **76.55**	40.67 | 65.84
Unknown | known w. ACC	73.61 | 89.35	80.72 | 92.04	**82.50 | 93.51**	69.04 | 88.55
Religious [train]				
UAS	74.99	79.76	79.52	**80.72**
LAS	61.61	67.94	65.94	**69.35**
ACC	83.31	87.62	85.91	**90.16**
Unknown | known w.	29.01 | 70.99	**29.50** | 70.50	26.58 | 73.42	14.07 | **85.93**
Different | shared lexicon	71.66 | 28.34	76.56 | 23.44	**85.05** | 14.95	43.47 | **56.53**
Unknown | known w. UAS	63.98 | 79.48	**70.17 | 83.77**	69.16 | 83.28	68.61 | 82.70
Unknown | known w. LAS	44.10 | 68.76	**53.31 | 74.06**	49.00 | 72.07	49.58 | 72.59
Unknown | known w. ACC	71.73 | 88.05	**81.80** | 90.06	75.87 | 89.55	75.87 | **92.50**

Table 7: Experiments Using Domain Based corpora

test sets do not stem from the same dialect, the shared lexicon is small. This could be due to the size of the SRCMF compared to contemporary language corpora, but more probably it is due to the heterogeneity of the texts in SRCMF, in particular concerning morpho-syntax and spelling, as shown in [3]. As a confirmation of the importance of shared lexicon: With an average sentence length of 7 words only, the *Normand* corpus should be easier to parse than the other dialects. Results do not clearly exhibit such differences (except for UAS), probably because the rate of known words is lower when evaluated on the *Normand* test set.

In any case, dialect turns out to be the most discriminative metadata among those evaluated, when it comes to predicting parsing results on Old French.

	Words	Units	Sentence	Sentence length
Champenois [train]	20005	3283	1821	10
Normand [train]	20028	3799	2639	7
Picard [train]	20011	3565	1538	13
Champenois [test]	21301	3440	1970	10
Normand [test]	13542	5503	1784	7
Picard [test]	40727	3042	3205	12

Table 8: Dialectal Corpora Characteristics

Train \ Test	**Champenois[test]**	**Normand[test]**	**Picard[test]**
Champenois [train]			
UAS	**86.07**	78.61	76.66
LAS	**76.30**	61.93	63.63
ACC	**93.41**	81.17	84.02
Unknown/known words	10.23 | **89.77**	**51.05** | 48.95	31.20 | 68.80
Different/shared lexicon	51.09 | **48.91**	**82.09** | 17.91	79.56 | 20.44
Unknown/known words UAS	**73.83 | 87.46**	72.83 | 84.63	66.38 | 81.32
Unknown/known words LAS	**59.14 | 78.25**	51.34 | 72.98	46.29 | 71.49
Unknown/known words ACC	**84.57 | 94.41**	72.59 | 90.12	67.99 | 91.30
Normand [train]			
UAS	74.54	**88**	73.77
LAS	59.31	**77.96**	60.48
ACC	81.12	**93.31**	82.55
Unknown/known words	34.14 | 65.86	11.25 | **88.75**	**38.77** | 61.23
Different/shared lexicon	82.24 | 17.76	43.90 | **56.10**	**87.05** | 12.95
Unknown/known words UAS	64.37 | 79.81	**78.53 | 89.20**	64.19 | 79.84
Unknown/known words LAS	45.30 | 66.58	**60.21 | 80.21**	46.86 | 69.11
Unknown/known words ACC	72.54 | 85.57	**82.01 | 94.74**	72.50 | 88.92
Picard [train]			
UAS	77.35	79.41	**85.14**
LAS	63.46	63.20	**75.90**
ACC	84.40	82.11	**93.25**
Unknown/known words	24.58 | 75.42	**45.57** | 54.23	11.16 | **88.84**
Different/shared lexicon	74.51 | 25.49	**82.42** | 17.58	60.03 | **39.97**
Unknown/known words UAS	66.15 | 81.00	**72.60** | 85.24	71.49 | **86.86**
Unknown/known words LAS	47.03 | 68.81	51.47 | 72.98	**55.29 | 78.49**
Unknown/known words ACC	75.34 | 87.34	72.93 | 89.78	**80.56 | 94.85**

Table 9: Experiments Using the Dialectal Corpus Segmentation

To go even further with this metadata, we make a complementary "leave one out" experiment using the dialect-based corpus segmentation, in order to determine which of its three distinct values is the most different from a machine learning point of view. The results in Table 10 show that the *normand* dialect seems to be the most remote one with the lowest amount of shared lexicon and known words. It leads to the lowest accuracy and LAS, however for both unknown and known words its UAS and LAS are higher than for the other two corpora. This can only be possible because it contains a higher rate of unknown words. This means that low proximity

between corpora does not necessary brings lower unknown words recognition.

The other 2 tested on	Champenois	Normand	Picard
UAS	84.18	**84.41**	82.06
LAS	**72.64**	70.41	70.98
ACC	**89.86**	86.56	88.54
unknown/known w.	17.04 ǀ **82.96**	**37.45** ǀ 62.55	25.05 ǀ 74.95
different/shared lex.	63.08 ǀ **36.92**	**75.97** ǀ 24.03	71.55 ǀ 28.45
unknown/known w. uas	75.25 ǀ 86.01	**79.16 ǀ 87.56**	73.56 ǀ 84.91
unknown/known w. las	58.82 ǀ 75.49	**59.84 ǀ 76.74**	56.39 ǀ 75.85
unknown/known w. acc	**84.29** ǀ 91.01	79.66 ǀ 90.69	76.97 ǀ **92.41**

Table 10: Leave One Out experiments Using the Dialectal Corpus Segmentation

4 Towards a General Methodology

This exploration based on metadata opens interesting perspectives. Given a new text with its associated metadata as an input, one can expect to develop a general methodology to find the best tagging or parsing model. Let the associated metadata be a set of attribute-value pairs. For instance, assume that the new input text has the following set of attribute-value pairs: *century=13th*, *domain=literary*, *form=verse*, *dialect=picard*. From the previous experimental results, one can find the best tagging/parsing model among those available. The selected parsing model for the input text would be the one leading to the best LAS among those associated with the text metadata. In our example, we have four possible candidate models, each one associated with an attribute-value pair:

- century=13th (LAS=80.75 trained on *13th century [train]*)
- domain=literary (LAS=73.63, trained on *literary [train]*)
- form=verse (LAS=71.40 trained on *verse [train]*)
- dialect=picard (LAS=75.90, trained on *picard [train]*)

In this example, the best parsing model for *century=13th* seems to be the one trained on *13th century [train]*, reaching a LAS of 80.75 (cf. Table 7).

This proposed methodology is still rudimentary. We are aware that, in our corpus, the metadata are correlated and interfere with each other. For instance, in SRCMF, texts written in *Picard* are all from the 13th century. This cannot be avoided due to the lack of available texts, let alone tagged corpora, of Old French. Furthermore, there is an unavoidable part of an arbitrary in the way metadata values are defined (the various distinguished domains, the time slicing, knowing that changes do not occurs specifically at the turn of two centuries...). The different sizes of the training sets are also another issue for such a general method.

5 Conclusion

In this paper, we have shown that machine learning could serve as a very effective corpus exploration strategy. Each experiment helps us to better understand the specificities of our highly heterogenous corpus. The originality of the approach we have followed is that it is focused on *metadata discrimination*. Machine learning engineering is usually more concerned with feature selection or parameter optimization, applied to stable training and test sets to gain better overall results. Here, the machine learning devices used are stable, but we vary the way training and test corpora are built, in order to evaluate their influence on the final result.

This work can be extended in various ways. With SRCMF, the metadata selection we have started out with should be investigated further, with the goal to build a complete decision tree for a given new text whose metadata are known. The ideal decision tree would provide the best possible labeled sub-corpus to use as a training set, to build the best possible model for this given text. To achieve this goal, the correlations between distinct metadata should also be investigated further. We are prudent concerning the generalizability of some of our conclusions, because the impacts of metadata are mixed with other factors: size of the available sub-corpora, lexical variation, effect of the combination of metadata... More experiments are necessary to clarify each of them.

Nevertheless, we believe that this global approach could be relevant in many contexts, as heterogenous corpora are increasingly becoming an important subject of parsing technologies.

References

[1] 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.

[2] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520, 2011.

[3] Gaël Guibon, Isabelle Tellier, Matthieu Constant, Sophie Prévost, and Kim Gerdes. Parsing poorly standardized language dependency on old french. In *13th Treebank and Language Theory (TLT)*, 2014.

[4] Céline Guillot, Alexei Lavrentiev, and Christiane Marchello-Nizia. Document 2. les corpus de français médiéval : Etat 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, July 2010.

[8] Francesco Mambrini and Marco Passarotti. Will a parser overtake achilles? first experiments on parsing the ancient greek dependency treebank. In *Eleventh International Workshop on Treebanks and Linguistic Theories*, pages 133–144. Edições Colibri, 2012.

[9] Alain Polguère and al. *Dependency in linguistic description*, volume 111. John Benjamins Publishing, 2009.

[10] 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.

[11] 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).

[12] Lucien Tesnière. *Eléments de syntaxe structurale*. Librairie C. Klincksieck, 1959.