

Umgeleitet von: <ilrstein@ling.uni-stuttgart.de>  
Von: lrec@lrec-conf.org  
Betreff: Your LREC 2014 Submission (Number 239)  
Datum: 24. Januar 2014 11:32:36 MEZ  
An: achim.stein@ling.uni-stuttgart.de  
Kopie: lrec@lrec-conf.org

Dear Achim Stein:

We are pleased to inform you that the following submission has been accepted for a Poster Session at the 9th International Conference on Language Resources and Evaluation:

Parsing Heterogeneous Corpora with a Rich Dependency  
Grammar

Please inform your co-author(s), if any. Each submission has been reviewed by three reviewers; please find attached their comments, if any.

Poster Sessions will be placed in parallel to Oral Sessions and therefore their length varies accordingly: during the session you will have the opportunity to describe your work and interact with interested conference participants. We request that you be at your Poster Session for the entire time slot. Further details for poster presenters will be detailed on the conference web site at due time.

The conditions for your participation in a Poster Session at LREC are the following:

1) you confirm your participation by registering for LREC by March 15, 2014 which is also the deadline for early registration with reduced fee (at least ONE author of the paper MUST register);

AND

2) you submit the final version of your paper by March 22, 2014 (FIRM deadline) in compliance with the "Authors'Kit", which is available on the LREC 2014 web site

(<http://lrec2014.lrec-conf.org/en/submission/authors-kit/>) and contains the Style sheet and the length for papers to appear in the Proceedings.

You can upload your final manuscript at the following site:

<https://www.softconf.com/lrec2014/main/>

You will be prompted to login to your START account and thus access to your User Console: from here please choose "Your current Submission(s)" for accessing the list of your submission(s). Click on this submission id for submitting the final full paper.

In case you need it, this is the passcode associated with your submission:

XXXXXXXXXXXX

We would like to remind you that there is no quality difference between Oral and Poster presentations, and there will be no difference between them in the Proceedings. The choice of style of presentation of your paper was made based on what would be the most informative and effective way to communicate the scientific contents of your presentation. There will be Chairpersons both for oral and poster sessions.

We remind you that you can update/revise the data for the LRE Map on the submission page and invite you to contribute to the "Share your LRs!" initiative as well.

Online registration and hotel booking, conference details and local information about Reykjavik will be soon available at the LREC2014 website (<http://lrec2014.lrec-conf.org/en/>).

If your organization requires you to have an official letter of acceptance from the LREC2014 Programme Committee in order to make arrangements to attend the conference, please request such a letter from the Conference Chair at: [lrec@ilc.cnr.it](mailto:lrec@ilc.cnr.it).

If you require an official letter of invitation from the local organizer (e.g. to obtain a visa), you have the possibility of requesting such a letter via the registration tool which will be available by mid-February on the Conference web site.

We would like to thank you for your participation in LREC and look forward to your presentation at the conference.

Best regards,  
LREC2014 Programme Committee

# Parsing Heterogeneous Corpora with a Rich Dependency Grammar

Achim Stein

Universität Stuttgart  
Institut für Linguistik/Romanistik  
achim.stein@ling.uni-stuttgart.de

## Abstract

Grammar models conceived for parsing purposes are often poorer than models that are motivated linguistically. We present a rich grammar model which is linguistically satisfactory and based on the principles of traditional dependency grammar. We show how a state-of-the-art dependency parser (*mate tools*) performs with this model, trained on the *Syntactic Reference Corpus of Medieval French* (SRCMF), a manually annotated corpus of medieval (Old French) texts. We focus on the problems caused by small and heterogeneous training sets typical for corpora of older periods. The result is the first dependency parser for Old French, and it is made publicly available.

**Keywords:** Dependency parsing, medieval texts, French

## 1. Introduction and Related Work

Our corpus is the *Syntactic Reference Corpus of Medieval French*, SRCMF (Prévost and Stein, 2013), and we describe the first parsing experiments for Old French (OF). For Modern French, the *French Treebank* (Abeillé and Barrier, 2004) has been converted to dependency structures, and parsing experiments reached an unlabelled attachment score of over 90% (Candito et al., 2010). The only other treebank for Old French is part of the corpus MCVF (Martineau, 2009), annotated according to the principles of the UPenn Treebanks.

The SRCMF differs from these treebanks in two respects: it is the first dependency-annotated corpus for Old French. The design of its grammar was driven by the desire to project a primarily linguistic analysis onto the corpus. The use of parsers is a secondary objective which did not matter for the original conception of the grammar model.

The SRCMF texts were annotated manually using the *Notabene* annotation tool (Mazziotta, 2010). *Notabene* stores the syntactic annotation as graphs in the Resource Description format (RDF). A high level of quality was ensured by an annotation process consisting of (a) two independent analyses by different annotators and (b) two independent reviews by the editors of the corpus. At both levels differences were discussed and resulted in a merged version.

In this contribution, our goal is not to improve parsers, let alone parsing algorithms. Rather, we consider parsing from a linguistic point of view and focus mainly on the following questions:

1. How does a parser perform with a rich grammar model not specifically designed to be “parser friendly”?
2. How does a parser perform with the relatively small amount of training data which is typical of high-quality manual syntactic annotation?
3. If a corpus contains heterogeneous texts (e.g. early OF verse as well as later OF prose), is it better to train separate models for each text type, or rather a general-purpose model?

Section 2. presents our corpus, the SRCMF. Section 3. is on methodology: it explains how the texts were annotated

and which tools were used. Section 4. describes three parsing experiments and interprets the results. Section 5. concludes.

## 2. The corpus

### 2.1. Texts

As of the time of writing, the SRCMF contains 15 texts with about 280 000 words. For this study, we selected the eleven texts listed in Table 1.<sup>1</sup> Some of their relevant properties will be mentioned in section 4. For more information see Stein and Prévost (2013) and the SRCMF website.<sup>2</sup> Totals for words and sentences refer to the actual training data used here. The three oldest texts were grouped into one sample, not for reasons of similarity (they are rather different), but because of their size. The texts marked with an asterisk (\*) are written in prose, the others are written in verse. The only mixed text (\*\*\*) is *Aucassin et Nicolette*, where prose and verse alternate, with a dominance of prose.

Text	Date	Words	Sent.
<i>Passion de Clermont</i>	late 10c.	(totals for	
<i>Vie Saint Legier</i>	late 10c.	these 3 texts)	
<i>Vie de Saint Alexis</i>	~ 1050	6642	803
<i>Chanson de Roland</i>	~ 1100	28643	3843
<i>Lapidaire en prose*</i>	mid 12c.	4699	467
<i>Tristan de Béroul</i>	end 12c.	26581	3237
<i>Yvain de Chretien de Troyes</i>	~ 1180	40529	3735
<i>Aucassin et Nicolette**</i>	~ 1200	9387	985
<i>Conqueste de Constantinople*</i>	> 1205	32960	2282
<i>Queste del saint Graal*</i>	~ 1220	39886	3049
<i>Miracles de G. de Coinci</i>	~ 1220	16996	1402
Total: 11 texts		206323	19803

Table 1: 11 texts of the SRCMF used in the parsing experiments (verse, prose\*, mixed\*\*)

<sup>1</sup>SRCMF pre-final version of February 2014. Minor corrections are in course. They will improve consistency and are expected to further improve the results published here.

<sup>2</sup><http://srcmf.org>.

## 2.2. Syntactic Properties and Grammar Model

“Old French” refers to a heterogeneous state of the French language. There is no variety which could be called OF “standard”, rather OF is a set of dialectal varieties with a large diachronic span, from the late 9c. to the middle of the 14c. OF is a null-subject language which often displays the verb in the second position (it is however unclear if these properties can be generalized). Word order is relatively free and adheres to information structural principles. Later OF gradually develops towards a more regular SVO word order while losing the distinction between nominative and oblique case. With respect to parsing, these syntactic properties make OF quite different from e.g. English or Modern French, and more similar to free word order languages with richer inflection, like German. The null-subject property is a further difficulty for parsing: the presence of a single argument of the verb does not imply the nature of this argument. In this respect, OF resembles Latin or Italian. The grammar model relies on the concept of dependency as defined by Tesnière (1965) and Polguère and Mel’čuk (2009). It uses a hierarchy of functions and structures to define the set of categories which are actually annotated in the corpus. They are listed in Table 2.

Tag	Function	Tag	Function
Apst	apostrophe	NgPrt	negative particle
AtObj	attribute of object	NMax*	non-maximum structure
AtSj	attribute of subject	NSnt*	non-sentence
Aux	auxiliation	Obj	object
AuxA	active auxiliation	Regim	oblique
AuxP	passive auxiliation	Rfc	reflexive clitic
Circ	adjunct	Rfx	reflexive pronoun
Insrt	comment clause	RelC	coordinating relator
Cmpl	complement	RelNC	non-coordinating relator
GpCoo*	coordinated group	SjImp	impersonal subject
Coo*	coordination	SjPer	personal subject
Intj	interjection	Snt*	sentence
ModA	attached modifier	VFin*	finite verb
ModD	detached modifier	VInf*	infinitival verb
Ng	negation	VPar*	participle verb

Table 2: Functions and structures\* of the SRCMF grammar model

The use of the categories is explained on-line in the SRCMF guidelines.<sup>3</sup> Important annotation principles are the following:

1. The top node of a sentence (Snt) is a finite verb (VFin) which does not depend on another verb. Hence there is no coordination between main clauses.
2. Each structure is governed by a lexical word (verb, noun, adjective, adverb).
3. Functional words (conjunctions, articles etc.) depend on lexical words. In this respect, the SRCMF grammar is similar to Stanford dependencies, but differs from models like the *Turin University Treebank* (Bosco, 2004), where functional categories nodes govern lexical heads.

<sup>3</sup><http://srcmf.org/fiches/index.html> (in French)

The SRCMF model is “rich” in the sense that, compared e.g. to the dependency version of the *French Treebank* (FTBdep), built by Candito et al. (2010), the SRCMF has more semantically motivated categories (see Table 2). Candito et al.’s categories seem to be more surface orientated: they distinguish e.g. object categories by their preposition (*de-obj*, *a-obj*). In the SRCMF prepositional phrases are distinguished on purely functional grounds: they are annotated as *Cmpl* (*complément*: indirect object) or as *Circ* (*circonstant*: adjunct). Since even human annotators struggle with borderline cases between both categories, it is normal that parsers also encounter problems: they are often visible in a higher difference between labelled and unlabelled attachment scores.

Both corpora have underspecified relations. For example, *obj* (the most frequent relation in the FTBdep) labels the relation between the verb and its direct object, as in *proposent une solution* ‘suggest a solution’, as well as between the preposition and the nominal head it governs, as in (1) between *par* and *attentes*. The relation between the participle *préoccupés* and its prepositional object is *dep*, as opposed to *Cmpl* used in the equivalent SRCMF structure (2) to mark the argument status of the phrase.<sup>4</sup> SRCMF, on the other hand, does not distinguish between determiners and other modifiers of nouns, using the underspecified label *ModA* for both.

- (1) préoccupés [*dep* par [[*det* les] *obj* attentes]]  
‘troubled by the expectations’
- (2) préoccupés [*Cmpl* [*RelNC* par] [*ModA* les] attentes]]

The examples also show that FTBdep favours a mixed approach with respect to the relation between head and function words: the preposition governs the noun, but the determiner depends on the noun. In the SRCMF, both preposition and determiner depend on the noun. In the light of the fact that the grammar models differ, a direct comparison between parsing results seems difficult.

In the following experiments, the full set of SRCMF categories was used, with the following modifications:

1. The difference between active and passive auxiliation (he has *come AuxA* vs *he was killed AuxP*) was not retained.
2. In the CoNLL format produced by the *Notabene* tool, coordination is annotated as a complex category, e.g. *coord1\_Obj*. Only the function (e.g. *Obj*) was retained. Hence coordination is marked by (a) dependence on the same word and (b) the presence of the relator (*RelC*) in the second structure.
3. The SRCMF grammar uses double references to a single form (*duplicata*) for contracted forms like *nel* (for *ne* ‘not’ + *le* ‘him’), and for relative pronouns like *qui* ‘who’ (relator (*RelNC*) and subject (*SjPer*). In CoNLL, these duplicate references form complex categories (*Obj\_RelNC*). Although they complicate the parser’s task, they are retained, because the information would not be recoverable.

<sup>4</sup>In other contexts, e.g. the object of a passive main verb, FTBdep uses *p-obj*.

### 3. Preparation and Method

#### 3.1. Annotation and Tools

The eleven texts (see Table 1) were exported to CoNLL 2009 format using the *Notabene* annotation tool (Mazzotta, 2010). The dependency relations were adapted as described in section 2.2. Part of speech (pos) annotation was added using the verified *Cattex* tagset<sup>5</sup> of the BFM database<sup>6</sup> (Guillot et al., 2007). *Cattex* tags define part of speech and subcategories, e.g. *ADJqua* (*adjectif qualificatif*), but no inflectional categories. Therefore, in the CoNLL format, the complete tag was kept in the pos column, the morph column remaining empty. Unverified lemmas were added using TreeTagger (Schmid, 1997) trained on the Old French *Nouveau Corpus d'Amsterdam* (Kunstmann and Stein, 2007).<sup>7</sup>

For the analysis, we used the *mate tools* package including Bohnet’s graph-based dependency parser (Bohnet, 2010; Björkelund et al., 2010).<sup>8</sup> The tagger was trained on the training set using 10-fold jackknifing. When training and evaluation were done on the same text or combination of texts, we used a 90%/10% split by selecting every 10th sentence. Thus, in a combination of texts, the evaluation part always reflected the heterogeneity of the whole corpus. For each analysis, we indicate the values explained in Table 3, except for the first experiment (“One on One”), where we leave out the values for lemma and label in Table 6 for reasons of space.

abbrev.	scores	meaning
lemma	lemma accuracy	for unverified lemmas
pos	part of speech	for verified <i>Cattex</i> tags
LAS	labelled attachment	dep. and label correct
UAS	unlabelled attachm.	dep. correct, label ignored
label	label accuracy	label correct, dep. ignored
Umatch	exact match	whole sentence
Lmatch	exact match	whole sentence incl. labels

Table 3: Syntactic scores and abbreviations

#### 3.2. Lemmatisation

Lemmatisation for Old French is only partial: about 70% of the graphemic forms have lemmas, and these lemmas can contain ambiguities when more than one lemma is possible. For example, the verb form *volt* is lemmatised as *valoir/voler/voloir* (‘be worth’, ‘fly’, ‘want’), i.e. three different verbs. Table 4 opposes results achieved on the lemmatised corpus with the unlemmatised one, where the word forms were reproduced in the lemma column of the CoNLL format.

Overall accuracies do not improve significantly with this imperfect TreeTagger lemmatisation. But the scores for some dependents of the verb are slightly better, especially for the direct object (*Obj*) and the distinction between indirect objects (*Cmpl*) and adjuncts (*Circ*). They certainly

<sup>5</sup>[http://bfm.ens-lyon.fr/article.php3?id\\_article=176](http://bfm.ens-lyon.fr/article.php3?id_article=176)

<sup>6</sup><http://bfm.ens-lyon.fr>

<sup>7</sup>The TreeTagger parameters are available on <http://srcmf.org>.

<sup>8</sup><http://code.google.com/p/mate-tools/>.

We used version `anna-3.3.jar`.

	imperfect lemmas	lemmas = forms
lemma	81.07	86.60
pos	93.89	93.98
UAS	89.61	89.56
LAS	82.41	81.99
label	86.29	85.98
Umatch	60.00	60.20
Lmatch	37.68	37.37
SjPer LAS	83.43	83.51
Obj LAS	72.21	69.93
Cmpl LAS	66.12	64.16
Circ LAS	74.59	73.84

Table 4: Lemmatised vs unlemmatised corpus

could be improved further if lemmatisation were of better quality. We decided to include lemmatisation in all our experiments. Note also that in a medieval corpus, a lemma is more than a generalisation over inflectional forms since it also subsumes graphemic variants.

#### 3.3. Checking Reliability and Size

Compared to corpora for modern languages, medieval text corpora tend to be too small for certain types of evaluation. We performed a 10-fold cross evaluation (a.) for the total of the texts, and (b.) for the 40.000 word text *Yvain*.

Training and evaluation were carried out on ten different 90/10 splits of both texts. Table 5 shows how the standard deviations for the 10 parts increase with decreasing corpus size.<sup>9</sup>

train/eval:	all 90/10			<i>Yvain</i> 90/10		
	$\mu$	$\sigma$	CV	$\mu$	$\sigma$	CV
lemma	81.24	0.64	0.01	80.91	0.95	0.01
pos	94.08	0.16	0.00	93.49	0.38	0.00
UAS	89.68	0.37	0.00	87.33	0.88	0.01
LAS	82.62	0.39	0.00	78.62	0.96	0.01
label	86.59	0.22	0.00	83.61	0.91	0.01
Umatch	60.84	0.76	0.01	52.72	1.48	0.03
Lmatch	39.88	1.12	0.03	30.68	1.62	0.05

Table 5: 10-fold cross evaluations: mean and standard deviation, for the complete corpus and a 40.000 words text

The mean values in the left column can be taken as our best results. They are quite encouraging, if we consider that OF has graphemic variation and that lemmatisation is of uncertain quality. At the time of writing we are not aware of any other Old French parsing experiment which could provide us with a baseline for comparison. But our UAS are not far below Candito et al.’s (2010) for Modern French, and both UAS and LAS are better than those for Latin (Passarotti and Dell’Orletta, 2010), although Latin has less graphemic variation.

#### 3.4. Refining the Questions

In order to answer the questions asked in the introduction, we conducted the three experiments described in the fol-

<sup>9</sup>Abbreviations used in tables:  $\mu$ : mean,  $\sigma$ : standard deviation, CV: coefficient of variation

evaluated on:	[1050]	Rol.	[Lap.]	Yvain	Trist.	Conq.	Mir.	[Auc.]	Quest.	statistics		
[trained on: "1050" ( <i>Passion de Clermont+Saint Legier+Saint Alexis</i> ): 6008 words]										$\mu$	$\sigma$	CV
pos	—	76.54	76.15	70.42	69.73	71.92	67.62	70.36	73.53	72.03	3.16	0.04
UAS	—	74.52	71.89	61.52	64.58	65.12	57.27	64.71	63.95	65.45	5.47	0.08
LAS	—	58.34	55.45	45.88	47.43	51.14	42.28	47.87	48.91	49.66	5.19	0.10
Umatch	—	39.80	30.88	22.02	32.54	21.49	18.54	36.08	20.82	27.77	8.04	0.29
Lmatch	—	13.27	8.08	7.57	11.24	8.23	6.02	12.85	7.49	9.34	2.72	0.29
trained on: <i>Chanson de Roland</i>										$\mu$	$\sigma$	CV
pos	74.48	—	83.25	73.42	74.18	75.73	69.93	75.12	80.36	75.81	4.16	0.05
UAS	<b>70.21</b>	—	<b>79.16</b>	64.48	68.33	67.80	59.71	67.73	71.03	68.56	5.57	0.08
LAS	<b>55.93</b>	—	<b>66.90</b>	49.71	52.20	55.07	45.39	52.93	57.67	54.48	6.32	0.12
Umatch	<b>36.10</b>	—	<b>42.28</b>	25.53	36.49	24.85	21.79	37.43	30.21	31.84	7.29	0.23
Lmatch	<b>17.29</b>	—	<b>19.95</b>	8.66	13.96	9.74	6.66	15.45	12.3	13.00	4.53	0.35
[trained on: <i>Lapidaire en prose</i> : 5250 words]										$\mu$	$\sigma$	CV
pos	59.29	65.57	—	57.90	55.89	64.36	56.77	59.62	62.92	60.29	3.59	0.06
UAS	56.81	65.56	—	50.06	53.79	58.27	45.75	54.87	54.59	54.96	5.82	0.11
LAS	37.98	46.48	—	33.74	34.15	41.85	30.53	37.05	39.04	37.60	5.02	0.13
Umatch	23.93	29.36	—	15.53	22.20	15.95	12.92	28.41	17.9	20.78	6.15	0.30
Lmatch	5.95	6.68	—	3.44	5.62	4.87	2.93	6.88	6.94	5.41	1.55	0.29
trained on: <i>Yvain de Chretien de Troyes</i>										$\mu$	$\sigma$	CV
pos	61.27	65.73	71.11	—	80.74	75.93	77.92	79.98	87.59	75.03	8.58	0.11
UAS	59.32	68.00	69.32	—	75.46	72.27	68.85	74.71	<b>80.80</b>	71.09	6.37	0.09
LAS	40.18	48.45	54.43	—	60.39	58.68	54.86	61.96	<b>70.06</b>	56.13	9.03	0.16
Umatch	23.37	32.07	29.69	—	42.45	26.70	27.73	44.87	36.67	32.94	7.71	0.23
Lmatch	7.75	7.90	11.16	—	17.15	12.01	11.17	21.53	18.53	13.40	5.08	0.38
trained on: <i>Tristan de Béroul</i>										$\mu$	$\sigma$	CV
pos	65.81	73.15	72.90	84.19	—	76.77	78.39	80.71	86.20	<b>77.27</b>	6.64	0.09
UAS	61.85	71.85	71.28	77.06	—	70.73	68.30	74.45	78.46	<b>71.75</b>	5.23	0.07
LAS	44.82	55.24	56.60	63.90	—	58.15	54.62	61.97	67.16	<b>57.81</b>	6.85	0.12
Umatch	26.83	39.21	32.07	37.49	—	26.53	28.21	43.74	36.12	<b>33.78</b>	6.36	0.19
Lmatch	10.51	13.86	13.06	16.23	—	12.68	12.52	20.52	20.43	<b>14.98</b>	3.75	0.25
trained on: <i>Conquete de Constantinople</i> (prose)										$\mu$	$\sigma$	CV
pos	54.74	62.05	66.01	68.10	68.00	—	67.35	76.05	77.04	67.42	7.17	0.11
UAS	53.33	61.93	67.87	61.58	63.02	—	58.64	70.27	70.53	63.40	5.95	0.09
LAS	34.77	42.00	49.47	45.44	45.98	—	42.60	56.32	57.98	46.82	7.66	0.16
Umatch	20.06	27.00	30.17	23.42	29.25	—	20.36	38.90	26.89	27.01	6.12	0.23
Lmatch	4.70	6.84	8.08	8.02	9.29	—	6.58	18.04	12.7	9.28	4.24	0.46
trained on: <i>Miracles</i>										$\mu$	$\sigma$	CV
pos	60.70	63.81	68.90	80.23	79.29	79.46	—	78.65	84.07	74.39	8.66	0.12
UAS	59.30	66.53	66.27	73.56	74.60	74.00	—	73.78	76.44	70.56	5.90	0.08
LAS	40.70	46.65	50.88	58.88	58.65	59.89	—	59.40	63.84	54.86	7.94	0.14
Umatch	23.65	30.97	28.50	32.46	41.69	29.72	—	42.28	32.97	32.78	6.37	0.19
Lmatch	7.88	7.94	10.21	12.73	17.58	12.51	—	20.18	15.85	13.11	4.48	0.34
[trained on: <i>Aucassin et Nicolette</i> (prose and verse): 8475 words]										$\mu$	$\sigma$	CV
pos	58.61	63.45	69.44	72.90	71.73	79.14	68.52	—	79.44	70.40	7.15	0.10
UAS	57.19	63.84	67.61	63.57	66.66	74.00	58.64	—	70.23	65.22	5.64	0.09
LAS	38.62	44.56	51.35	48.04	49.12	61.88	42.76	—	57.16	49.19	7.61	0.15
Umatch	21.58	28.73	29.45	23.87	32.78	30.31	20.36	—	25.87	26.62	4.42	0.17
Lmatch	6.22	6.46	9.03	7.26	10.67	15.03	6.74	—	12.62	9.25	3.25	0.35
trained on: <i>Queste du Graal</i> (prose)										$\mu$	$\sigma$	CV
pos	61.19	65.08	69.91	80.17	77.83	78.91	75.75	80.64	—	73.69	7.40	0.10
UAS	58.59	66.87	69.23	72.79	72.98	76.20	65.94	75.92	—	69.82	5.94	0.09
LAS	40.51	46.67	54.32	58.24	56.38	64.49	51.29	63.06	—	54.37	8.09	0.15
Umatch	24.34	32.23	31.12	33.29	39.59	30.06	26.70	47.58	—	33.11	7.40	0.22
Lmatch	7.75	9.46	11.88	13.62	14.05	14.61	9.98	25.93	—	13.41	5.61	0.42

Table 6: Cross evaluation on single texts of roughly equal size (except for three smaller texts: "1050", "Auc.", "Lap.")

trained on:	all the texts except the evaluated text									statistics		
evaluated on:	-1050	Rol.	Lapid.	Yvain	Trist.	Conq.	Mir.	Auc.	Quest.	$\mu$	$\sigma$	CV
lemma	71.89	67.70	80.44	75.98	75.50	74.38	76.54	39.88	84.48	71.87	12.90	0.18
pos	79.55	87.90	87.19	90.34	89.39	89.44	87.94	90.57	94.48	88.53	3.99	0.05
UAS	74.83	85.26	83.85	85.35	85.13	85.47	79.32	87.11	90.06	84.04	4.46	0.05
LAS	62.83	73.69	74.89	75.95	74.74	76.94	69.22	78.76	82.88	74.43	5.72	0.08
label	69.98	77.96	80.93	81.39	79.71	82.83	77.44	83.50	87.06	80.09	4.80	0.06
Umatch	41.59	57.25	53.32	48.86	57.46	43.21	39.02	62.54	55.07	50.92	8.16	0.16
Lmatch	22.79	30.86	32.55	27.15	31.85	24.67	21.26	37.56	34.14	29.20	5.52	0.19

Table 7: Cross evaluation “leave one out”: training on all the texts except one, evaluation on that one text.

lowing sections. They simulate the real-world scenarios of a linguist wanting to annotate a new Old French text. Considering the heterogeneity of OF, he or she will face the question if a general parser model suffices, or if several specific models have to be trained. We try to answer the following questions:

1. **Single-text models.** Given the heterogeneity of Old French, how does a model trained on a given text perform on each of the other texts? This question will be studied in the “one-on-one” cross evaluation (section 4.1.).
2. **Annotate a new text.** If the maximum of available data is used for training, would the annotation results depend on the type of text? Is it possible to isolate specific properties of the texts? These questions will be studied in the “leave-one-out” cross evaluation (section 4.2.).
3. **Prose or verse.** Does it matter if the new text is prose or verse? Most syntacticians would say it does. But this does not necessarily mean that a dependency model performs differently on both text types. This question will be studied in the prose-verse cross evaluation (section 4.3.).

## 4. Experiments

### 4.1. Cross Evaluation “One on One”

In this cross evaluation, the parser is trained on one specific text of the SRCMF, and evaluated on each of the other texts. An equal training set size of about 16.000 words was defined by the size of the smallest text considered to be of sufficient size to provide us with reliable results, *Miracles de G. de Coinci*. For the sake of comparison, we also included the three smaller text samples (the three oldest texts “-1050”, *Lapidaire* and *Aucassin e Nicolete*) in the cross analysis, their sizes are indicated in Table 6, but the results in these columns should be treated with greater care.

As expected, the results are worse than the results obtained from the combined training corpus (left column of Table 5), but our goal is to make global statements about the resemblance between individual texts. We comment on some results highlighted in bold font in Table 6:

1. The *Roland* model does very well for the oldest texts. It achieves the highest attachment scores and sentence matches in *Lapidaire* (compare horizontally), and the

three other 11th century texts (*1050*) were analysed better by this model than by any of the others (compare vertically).

2. The *Tristan* model gets the best mean values (column  $\mu$ ) for all scores, whereas *Lapidaire*, probably due to its limited size, gets the worst values.
3. *Queste* is the text which is analysed best for several models: the UAS and LAS of the *Yvain* model evaluated on *Queste* are the best results of the table.
4. *Miracles de Nostre Dame* seems to be quite different from the other texts: all the models achieve scores below average on this text. The mean attachment scores across all models (not given in Table 7) for *Miracles* are 52.01 (UAS) and 41.94 (LAS). Some reasons for these exceptionally poor results will be given in the discussion of the following experiment.
5. The model trained on the prose text *Conqueste* has mean attachment scores of 63.40/46.82 (UAS/LAS). The model trained on the other prose text, *Queste*, does better with 69.82/54.37. This difference may be due to the text genre: *Conqueste* is a chronicle; the sentences are long, but not complex, preferring enumerations to syntactic embeddings.
6. Surprisingly, the model trained on the only mixed text, *Aucassin et Nicolete*, also performed quite poorly, although one might expect better results on prose and verse from a text which has both. The dominance of the prose part is probably too high. We will come back to the prose/verse opposition in section 4.3.

### 4.2. Cross Evaluation “Leave One Out”

In this evaluation, the parser was trained on all the texts except the one to be evaluated. The result can be an indicator for the situation where a new text is analysed, e.g. to enrich the corpus. We suppose that results will be similar to those achieved on the text which is most similar to the new text. The results are presented in Table 7.

1. By far the best attachment scores are attained on the *Queste du Graal*. They are close to the results achieved on the complete 90/10 split corpus (Table 5, left column).
2. With the other eight texts, scores are worse. The lowest attachment scores are attained for the three oldest



texts (column “-1050”), which might indicate the diachronic limit of a globally trained model. An interesting question which could be answered by a more detailed experiment bearing on the oldest French texts would be the impact of part of speech tagging accuracy: it is possible that the parser is hampered by the fact that the tagger does not cope with the less “conventionalised” orthography in this particular period.

3. The second worst result was achieved on the *Miracles de Nostre Dame*. An in-depth investigation of the reasons is beyond the scope of this article, but some indications are given by Rainsford et al. (2012). This text contains a considerable part of lyric poetry, a genre not present in other texts of the SRCMF, and the syntax is peculiar, with significantly more elements appearing pre-verbally than in the other texts: 33,3% of the structures have three or more pre-verbal elements, 17 structures even have five or more pre-verbal elements. The authors also note that some orders of pre-verbal elements are typical of verse, e.g. the subject preceding an adverbial (*SjPer-Circ-V*).

### 4.3. Cross Evaluation for Prose and Verse

To answer our third question, we created two subcorpora (Table 8). The first is composed of two texts in prose, the second of three texts in verse. The subcorpora are equal in word size. The prose texts have a higher word per sentence ratio of 13.7 (compared to 10 for verse). On the diachronic scale, both subcorpora are comparable: the manuscript dates span a period of no more than 40 years.

	texts	words	sent.
prose	<i>Conqueste + Queste</i>	72846	5331
verse	<i>Miracles + Tristan + Yvain</i>	72854	7320

Table 8: subcorpora for prose and verse

The parser was trained on 90% of both subcorpora, and evaluated on (a) the complete other subcorpus and (b) 10% of the same subcorpus, for comparison.

trained on:	verse90		prose90	
	prose	verse10	verse	prose10
lemma	80.36	83.51	70.10	88.53
pos	89.10	92.85	82.32	96.63
UAS	84.35	86.17	74.93	92.63
LAS	74.98	77.50	62.01	86.75
label	81.22	82.68	69.95	89.66
Umatch	42.31	56.01	39.13	56.85
Lmatch	24.16	34.29	17.30	36.21

Table 9: Cross evaluation prose/verse

Parsing prose with a model trained on verse leads to significantly better results than the contrary, with differences of more than 10% for label accuracies and LAS. Considering the corpus size, the results are not much worse than those for global training in Table 5. Why does the parser, when trained on prose, perform so badly for verse? Even without going into details, it is definitely true that verse syntax

is more variable, especially in a language with rather free word order like OF. A parser trained on verse has seen more of these variants, and hence performs better. The relatively good results (compared to the global result in Table 5) obtained on the prose 90/10 split are another indicator for the relative uniformity of prose syntax.

### 4.4. General Observations

The difference between labelled and unlabelled attachment scores is quite high (about 10%, sometimes even higher). This means that many words are attached correctly, but the parser has difficulties to guess the right label. This is probably the price we have to pay for our “rich” annotation model: we had mentioned the fact that several oppositions between categories are semantically motivated and therefore hard to learn for a parser.

The generally low score of exact matches could indicate that a high number of dependencies are easy to learn, whereas a small number of dependencies are difficult to learn. The variation of exact matches is about three times higher than the variation of the attachment scores, as the coefficients of variation (CV) in Tables 6 and 7 indicate.

## 5. Conclusions and Future Work

We described the first dependency parser for Old French. It was trained on a 200 000 word subset of the SRCMF dependency treebank, using the *mate tools*. We showed that even on a corpus with graphemic variation and poor lemmatisation, good parsing results can be achieved with relatively few training data and a “rich”, i.e. semantically expressive grammar model. We carried out experiments in order to show that a heterogeneous corpus with texts of different types and from different historical subperiods (between the late 10c. and the early 13c.) can provide a general-purpose model, so that even for a multi-variety language like OF a single parser model may suffice.

On a 90/10 training/evaluation split of eleven OF texts, the *mate* parser attained a UAS of 89.68% and a LAS of 82.62%.<sup>10</sup> There is no baseline for Old French, but these results are quite satisfactory if the graphemic variability and the incomplete lemmatisation of the texts are taken into account. We carried out three cross evaluation experiments to show in which ways heterogeneity, which is typical of medieval corpora, influences the parsing results.

In addition to the text-specific “one-on-one” cross evaluation (experiment 1), we tried to circumvent the sparse data problem by carrying out a “leave-one-out” cross evaluation (experiment 2). We also showed that training on verse provides us with a better general-purpose model than training on prose, probably because the parser encounters more syntactic variation in verse texts (experiment 3).

## 6. References

Anne Abeillé and Nicolas Barrier. 2004. Enriching a French treebank. In *4th international conference on language resources and evaluation LREC, Lisbon*.

<sup>10</sup>These models for *mate tools* will be publicly available on <http://srcmf.org>.

- Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August. Coling 2010 Organizing Committee.
- Cristina Bosco. 2004. *A Grammatical Relation System for Treebank Annotation*. PhD Thesis, Università degli Studi di Torino.
- Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical French dependency parsing: treebank conversion and first results. In *Proceedings of LREC'2010, La Valletta, Malta*.
- Céline Guillot, Christiane Marchello-Nizia, and Alexej Lavrentiev. 2007. La base de français médiéval (BFM): états et perspectives. In Pierre Kunstmann and Achim Stein, editors, *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*. Steiner, Stuttgart.
- Pierre Kunstmann and Achim Stein. 2007. Le Nouveau Corpus d'Amsterdam. In Pierre Kunstmann and Achim Stein, editors, *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*, pages 9–27. Steiner, Stuttgart.
- France Martineau, editor. 2009. *Le corpus MCVF. Modéliser le changement: les voies du français*. Université d'Ottawa, Ottawa.
- Nicolas Mazziotta. 2010. Building the 'Syntactic Reference Corpus of Medieval French' using notabene RDF annotation tool. In *Proceedings of the 4th Linguistic Annotation Workshop (LAW IV)*.
- Marco Passarotti and Felice Dell'Orletta. 2010. Improvements in parsing the Index Thomisticus treebank. revision, combination and a feature model for medieval latin. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Alain Polguère and Igor Mel'čuk, editors. 2009. *Dependency in Linguistic Description*. Benjamins, Amsterdam, Philadelphia.
- Sophie Prévost and Achim Stein, editors. 2013. *Syntactic Reference Corpus of Medieval French (SRCMF)*. ENS de Lyon; Lattice, Paris; Universität Stuttgart, Lyon/Stuttgart, <http://srcmf.org>.
- Thomas Rainsford, Céline Guillot, Alexei Lavrentiev, and Sophie Prévost. 2012. La zone préverbale en ancien français : apport des corpus annotés. In *Actes du 3e Congrès Actes du 3e Congrès Mondial de Linguistique Française (CMLF), Lyon, July 2012*. Institut de Linguistique française, Paris.
- Helmut Schmid. 1997. Probabilistic part-of-speech tagging using decision trees. In Daniel Jones and Harold Somers, editors, *New Methods in Language Processing*, Studies in Computational Linguistics, pages 154–164. UCL Press, London, GB.
- Achim Stein and Sophie Prévost. 2013. Syntactic annotation of medieval texts: the syntactic reference corpus of medieval French (SRCMF). In Paul Bennett, Martin Durrell, Silke Scheible, and Richard Whitt, editors, *New Methods in Historical Corpora*, Corpus Linguistics and International Perspectives on Language, CLIP Vol. 3, pages 275–282. Narr, Tübingen.
- Lucien Tesnière. 1965. *Éléments de syntaxe structurale*. Klincksieck, Paris, 2 edition.