Old French Dependency Parsing: Results of Two Parsers Analysed from a Linguistic Point of View

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

Abstract

The treatment of medieval texts is a particular challenge for parsers. I compare how two dependency parsers, one graph-based, the other transition-based, perform on Old French, facing some typical problems of medieval texts: graphical variation, relatively free word order, and syntactic variation of several parameters over a diachronic period of about 300 years. Both parsers were trained and evaluated on the *Syntactic Reference Corpus of Medieval French* (SRCMF), a manually annotated dependency treebank. I discuss the relation between types of parsers and types of language, as well as the differences of the analyses from a linguistic point of view.

Keywords: dependency parsing, medieval texts, Old French, error analysis

1. Introduction

1.1. Previous and related work

The treatment of Medieval texts in general, and Old French (OF) in particular is a challenge for NLP.

The first issue is variation. Previously I have shown that the problem of syntactic variation in a treebank of OF texts spanning over 300 years is not unsurmountable, since a global model trained on the totality of the texts is almost as accurate as specific models trained for subcorpora (Stein, 2014). The results of the graph-based *mate tools* parser (Bohnet, 2010; Björkelund et al., 2010) were quite satisfactory (LAS: 82.62%).

The second issue is the relatively free word order in OF, in combination with the null-subject property, i.e. sentences need not have an overt subject (like e.g. Modern Italian or Spanish).

The third issue is that with regard to inflection, OF is closer to e.g. Modern German than to Modern English. Verbs are marked for person, number, and tense/mood; nouns, pronouns, and adjectives are marked for number and case (OF has a two-case system), see also Table 3.

1. Issue 1 is not the main focus of this paper, but compared to the previous experiment (Stein, 2014), a different part-of-speech tagger was selected in order to improve the morphological analysis despite of the orthographical variation (see section 3.2.).

2. With regard to issue 2, it has been argued that transition-based parsers are assumed to be more suitable for languages with relatively free word order. This is due to the fact that transition-based parsers are more "dynamic", in that they learn sequences of transitions that are applied successively to the words of a sentence by taking words from a queue, and adding processed words to a stack. Graph-based parsers are more "static", in that they learn a model over complete dependency graphs by summing up all the attachment scores in a sentence.

3. With regard to issue 3, "joint morphological and syntactic disambiguation is especially important for richly inflected languages, where there is considerable interaction between morphology and syntax such that neither can be fully disambiguated without considering the other" (Bohnet et al., 2013). Further properties of graph-based vs transition-based approaches are discussed by Bohnet and Kuhn (2012). These arguments suggest that a transition-based parsing and joint morphological and syntactic analysis might improve the parsing results for Old French.

The *Joint Transition-based Parser* (Bohnet et al., 2013), henceforth JTP, was selected to verify this hypothesis. This parser provides the following technology: transition-based dependency parser, beam-search and early update, graph-based completion model, joint Part-of-Speech tagging, joint Morphologic tagging, Hash-Kernel.[1]

1.2. Goals

In this paper, I evaluate both parsers on an Old French treebank. The experiments were carried out on the *Syntactic Reference Corpus of Medieval French (SRCMF)* (Prévost and Stein, 2013). Work based on a previous version of this corpus has shown that (a) with the graph-based dependency parser good parsing results can be achieved even with the limited amount training data that is available for Old French and with a relatively rich, i.e. linguistically satisfactory grammar model, and (b) that even for a heterogeneous corpus consisting of different text types and spanning three centuries, a general model can be trained on the complete data, rather than training text-type specific models, see Stein (2014).

Contrary to other experiments the parsing results will also be evaluated from a philological and linguistic point of view. Since the goal is not to improve parsers or parsing algorithms, no effort was made to adapt the annotation to the weak (or strong) spots of a given parser or parsing technique. The focus will be on the following questions:

1. Since Old French is a language with relatively free word order, does the joint transition-based parser achieve better results than the graph-based parser?
2. Do particular syntactic properties of the language matter for the choice of the parser?

[1] From the desciption on the *mate tools* website: https://code.google.com/archive/p/mate-tools/wikis/ParserAndModels.wiki

3. How do the differences in the global scores relate to differences in the analyses of particular grammatical functions? In other words: is one of the parsers better in every respect, or are the improvements limited to a better performance with particular functions (and therefore dependent on the frequency of these)?

Section (2.) presents the *Syntactic Reference Corpus of Medieval French* (SRCMF). The parsing experiments with both parsers are described in section (3.). Section (4.) provides a linguistic assessment of the differences between the predictions of both parsers, sentenced-based as well as with regard to particular categories. Section (5.) concludes.

2. The corpus

2.1. Texts

The SRCMF contains the texts listed in Table 1. I limit myself to the discussion of the properties which matter most for the comparison of parsing results. For a more detailed introduction to SRCMF see Stein and Prévost (2013) and the corpus website.[2] The texts marked with an asterisk (*) were not used in the experiment.

Title	Date	Words
**Serments de Strasbourg*	842	115
**Sequence de sainte Eulalie*	881	189
**Passion de Clermont*	950-1000	2842
Vie Saint Legier	950-1000	1388
Vie de saint Alexis	around 1050	4868
Chanson de Roland	around 1100	28997
Lapidaire en prose	middle of 12c.	4765
Yvain de Chretien de Troyes	1177-81	41702
Quatre Livres des Rois	end of 12c.	13061
Tristan de Beroul	end of 12c.	27052
Conqueste de Constantinople	after 1205	33969
Queste del Saint Graal	around 1220	40636
Miracles de G. de Coinci	1218-1227	22418
Roman de la Rose de J.de Meun	1269-1278	19462
Aucassin et Nicolete	around 1300	9946

Table 1: SRCMF 0.9: texts, dates, word count

The CoNLL export version of the texts contains 242 946 word tokens and 23 818 types. Punctuation was not present (modern punctuation appears only in modern transcriptions). Orthographical variation is considerable: the type-token ratio is more than twice as high (0.099) than in Modern French texts (0.048). This has obvious negative consequences for the precision of part-of-speech taggers.

2.2. Syntactic properties and the 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 early 14c. OF is a null-subject language and often has 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. Modern English and more similar to free word order languages with richer inflection like German.

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. Minor modifications were applied to the CoNLL version used for this paper. The list of categories is in Table 2.

Abbrev.	Function	mod. Example/Explanation
Apst	apostrophe	*Sire* 'Lord!'
AtObj	attribute of object	*on nomma Paul roi*
AtSj	attribute of subject	*Paul était roi*
Aux	auxiliation	non-finite verb forms
Circ	adjunct	all kinds of adverb(ial)s
Cmpl	prep. object	indirect/locative arguments
Ignorer	forms to ignore	e.g. errors in the manuscript
Insrt	comment clause	Dame *fait-il...* 'says he'
Intj	interjection	*Ha sire, fait Galaad*
ModA	attached modifier	lexical or clausal modifiers
ModD	detached modifier	dislocated structures
Ng	negation	*ne* (first part of negation)
NgPrt	negative particle	*pas* etc. (second part of neg.)
Obj	direct object	*Paul voit le roi*
Regim	oblique	infinitival clauses
RelC	coordinating relator	conjunctions, e.g. *et*
RelNC	non-coord. relator	conjunctions, rel. pronouns
Rfc	reflexive clitic	*il se casse* 'it RFL breaks'
Rfx	reflexive pronoun	*soi-même* 'himself'
SjImp	impersonal subject	see example (3)
SjPer	personal subject	*Paul voit le roi*

Table 2: SRCMF categories (CoNLL version)

In addition, the corpus contains composed functions for agglutinations, e.g. *Obj_Ng* for *nel* < *ne+le* 'not'+'it'. The most important annotation principles are the following:

1. The root node of a sentence is a finite verb which does not depend on another verb.
2. Non-finite verbs depend on the finite verb (e.g. auxiliaries govern participles, the relation is *Aux*).
3. Arguments of the verbs depend on the finite verb.
4. Each structure is governed by a lexical word (verb, noun, adjective, adverb).
5. Functional words (conjunctions, articles etc.) depend on lexical words, i.e. articles on nouns, or conjunctions and relative pronouns on subordinate verbs.

The use of the categories is explained in detail on-line in the SRCMF guidelines.[3]

The texts were manually annotated using the *Notabene* annotation tool (Mazziotta, 2010). High-quality annotation was ensured by an annotation process consisting of (1.) two independent analyses by different annotators and (2.) two independent reviews by the principal investigators. At both levels differences were discussed, and resulted in a merged version.

[2] http://srcmf.org.

[3] http://srcmf.org/fiches/index.html (in French)

3. The parsing experiment

3.1. Preparation of the texts

For the experiment, twelve texts from SRCMF were selected, i.e. all the texts in the Treebank except for the texts marked with an asterisk in Table 1. The problem of diachronic and text-type heterogeneity was discussed in Stein (2014).

The texts were annotated and exported to CoNLL 2009 format using the conversion of the *Notabene* annotation tool. The format reproduces the complete SRCMF annotation model, but simplifies coordinating structures. In SRCMF, coordinated elements are attached to a coordination node. Since CoNLL does not support empty nodes, the first conjunct governs the others. The dependency is labelled *coord-X*, where X is the syntactic category, e.g. *coord-Obj*. These hybrid categories were reduced to the function part (i.e. *coord* was omitted) in the experiments described here. This decision was motivated by the fact that OF also allows coordination of different functions, and that users would not be able to recover these functions if omitted. Part of speech (pos) annotation was added using the verified *Cattex* tags[4] of the BFM database[5] (Guillot et al., 2007). These pos tags indicate part of speech and eventual subcategories (e.g. *ADJqua* for 'qualitative adjective'), but no morphological features.

In addition to the verified dependencies and pos tags (as contained in the SRCMF distribution), morphological features and lemmas were added automatically, using *TreeTagger* (Schmid, 1997) with parameters trained on the *Nouveau Corpus d'Amsterdam* (Kunstmann and Stein, 2007), an Old French 3 Mio word corpus with manual morphological annotation. From this tagset, the features for gender, number, case, and person were added to the CoNLL 'feature' column, while leaving the *Cattex* tags untouched (the features are listed in Table 3). The *TreeTagger* was used to add a lemma string containing one or more possible lemmas, e.g. `estre|fuir` for the ambiguous form *fui*, to 94% of the forms (Stein, 2007). These lemmas are not verified. The data contains 6318 lemma types (different lemma strings).

3.2. Training

For each of the experiments described below, I used 90:10 splits and 10-fold cross-validation, with 20 771 sentences in the training data, and 2 307 sentences for evaluation.

The graph-based dependency parser was the *mate tools* parser (Bohnet, 2010; Björkelund et al., 2010).[6] It was compared to the more recent *Joint Transition-based parser* (Bohnet et al., 2013).[7], which performs a joint analysis of the morphological and syntactical levels. The authors consider it to be "the first joint system that performs labeled dependency parsing" as well as "the first joint system that achieves state-of-the-art accuracy for non-projective dependency parsing" (*ibid*, p.1456). For Old French the transition-based approach was particularly promising because the algorithm specifies an interaction between the analyses of the different annotation layers, i.e. between part of speech, morphological features, and dependencies. The hypothesis is therefore that the transition-based joint parser would improve the results medieval texts in general and for Old French in particular. It is based on the observations and experiments for a number of languages described in Bohnet et al. (2013).

Graphical variation is a general feature of medieval texts, regardless of the particular language. It makes the assignment of a correct part of speech tag (and subsequently lemmatisation) more difficult. It is true that pos tagging scores can to some extent be improved using generalised graphical forms, e.g. based on graphemic rules (cf. for example Souvay and Pierrel (2010) for Middle French), but from a philological perspective the original word form is important and often the result of a deliberate choice on behalf of the transcriber or editor of the manuscript. In the particular case of Old French, covering a time span more than three centuries, graphical normalisation is even more complicated, since rules would have to be sensitive not only to particular periods but also to particular regions. Therefore no normalisation was applied, and all tools were trained on the original graphical form.

Traditional dependency parsers like the graph-based parser are part of a pipeline where part of speech assignment (and eventually lemmatisation) is prior to the dependency parsing. Hence they are sensitive to inaccurate part of speech assignment. As mentioned above, Old French has a relatively rich inflection compared to e.g. Modern English or French: it has a gender-specific two-case system marked on nominals and adjectives, as well as person and tense marked on verbs. I assume that what Bohnet et al. claim for other case-marking languages (see the quotation in section 1.1.) also holds for Old French. I therefore expect the joint transition-based parser to improve the pos tagging results *and* the dependency analysis.

	case	gender	number	person
verb	–	–	+	+
noun	+	+	+	–
adjective	+	+	+	–
determiner	+	+	+	–
pronoun	+	+	+	+
# of values	2	2	2	3

Table 3: Use and values of morphological features

Finally, it is well known that the accuracy of the part of speech tagging has considerable influence on the parsing accuracy. The *mate tools* contain a pos tagger, but it performs less well than other state-of-the-art taggers. Since the goal here is to compare the parsers on a similar basis, the *mate tools* tagger was replaced with *Marmot* (Müller et al., 2013). This improved the accuracy of pos tagging (from 94.77% to 95.49%) and feature tagging (from 91.44% to 93.72%) to a score which is on a par with the joint transition-based parser (95.78% for tags, 93.21% for features).

[4] http://bfm.ens-lyon.fr/article.php3?id_article=176

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

[6] http://code.google.com/p/mate-tools/

[7] The parser is available at https://code.google.com/p/mate-tools/wiki/ParserAndModels.

4. Evaluation

The evaluation file (gold) has 2307 sentences with 24090 tokens (10.4 tokens per sentence).
The abbreviations "GPM" for the graph parser with the *Marmot* tagger, and "JTP" for the joint transition-based parser will be used.

4.1. General evaluation

In Table 4, the direct comparison of both parsers shows for the joint transition parser a slight advantage of +0.78 percent points over the result attained with the graph parser in the labelled attachment score (LAS). The more noticeable difference is the score of exact matches at sentence level: the JTP outperforms the GTP by +5.76 points. This means that the errors of the JTP occur in fewer sentences. The average of incorrect labelled attachments per false sentence is 1.87 (GPM) and 1.97 (JTP) respectively.

	graph GPM	joint trans. JTP	diff.
pos acc.	95.49%	95.78%	+0.29
feat acc.	93.72%	93.21%	–0.51
UAS	91.54%	91.75%	+0.21
LAS	85.18%	85.96%	+0.78
label acc.	88.51%	89.06%	+0.55
ex. match (UAS)	64.02%	66.67%	+2.65
ex. match (LAS)	41.83%	47.59%	+5.76

Table 4: Scores of graph parser and joint transition parser

These differences confirm the hypothesis that transition-based parsers perform better on highly inflecting and free word order languages, and add another language to Bohnet et al.'s list of results for this type of languages.

4.2. Sentence-based comparison

Kübler et al. (2009, 80) observe that sentence-based scores can be a meaningful complement of word-based evaluation. Here, the difference between the "exact match" scores in Table 4 is analysed in greater detail, using again calculated scores, but on a per-sentence basis. The following observations are related to the figures in Table 5, where "T" means 'true prediction' and "F" means 'false prediction'.

- The first line reproduces the exact LAS match of Table 4. Line 2 shows that most of these matches are correctly predicted by both parsers (Table 5, line 3). For a more detailed account on per-category parser agreement see section 4.4.
- Sentence length: the average token per sentence length with correct LAS is 6.6 for the GPM and 6.9 for the JTP (line 4).
- With respect to correctly tagged sentences, the JTP is about 1.5 point ahead of the GPM. Just like in the case of LAS, errors seem to be slightly more concentrated (line 5).
- It has been mentioned that parsing accuracy highly depends on correct pos tagging. Therefore, when only sentences with tagging errors are evaluated, LAS accuracy is low, but the JTP performs slightly better (line 6).

in sentences	GPM	JTP
LA=T	965 (41.8%)	1098 (47.6%)
LA=T both parsers	835 (36.2%)	
LA=T for this parser	130 (5.6%)	263 (11.4%)
ø tokens in LA=T	6.6	6.9
pos=T	1611 (69.83%)	1646 (71.35%)
pos=F & LA=T	97 (4.2%)	116 (5.0%)

Table 5: Sentence-based results (T=true, F=false, LA=labelled attachment, pos=part of speech)

4.3. Linguistic discussion of selected relations

In this section, some significant LAS differences for specific categories related to particular syntactic phenomena will be addressed, as well as category-specific differences between precision and recall, i.e. "the percentage of dependencies with a specific type in the parser output that were correct" vs "the correct percentage of dependencies with a specific type in the test set that were correctly parsed" (Kübler et al., 2009, 79).

4.3.1. Left dislocation

ModD, the "detached modifier", marks the relation between a dislocated structure and its governor. The dislocated structure is often a noun phrase or a relative clause. It depends on a resumptive pronoun, e.g. the locative (*en*) in (1) or the subject (*il*) in (2):

(1) [Des helmes]$_i$ clers li fuus en$_i$ escarbunet
of the helmets bright the fire of them shines
From the bright helmets shone the light.

(2) [Rex Chielperings]$_i$ il$_i$ se fud mors
King C. *he* REFL *was died*
King Chielpering died.

Thus *ModD* normally occurs with long-distance dependencies. Here both parsers have the same low recall: they only predict 4 of 35 (11.35%) *ModD* correctly. However, the precision score of the GPM (22.2%) is affected by 14 incorrect predictions, compared to only 4 of the JTP (50%). *ModD* is a good example for a category which is linguistically relevant: dislocation and the development of clitics are much discussed topics in Romance diachronic linguistics. The results also seem to indicate that a rigid implementation of dependency can be a problem for parsers: from a theoretical point of view, it is probably correct to assume a dependency relation between the dislocated structure and the resumptive prounous, as does the SRCMF grammar.
Contrary to this theory-driven analysis, in most other dependency treebanks (or in conversions from constituent models) these elements are on the same level. An example is the Danish Dependency Treebank, where e.g. a dislocated subject is attached to the main verb (with *xtop*), on the same level as the resumptive subject pronoun (*subj*), as

shown in Kromann and Mikkelsen (2003, 219); the relation between the two is marked by a secondary edge (Fig. 1).
It is possible that the more deeply embedded relation *ModD* in the SRCMF grammar is more difficult to learn than a generalised relation depending directly on the verb. The

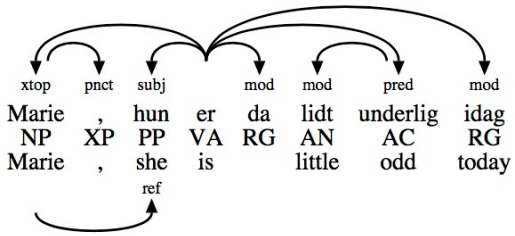


Figure 1: Left dislocation in the Danish Dependency Treebank (Kromann and Mikkelsen, 2003, 219)

figures show the SRCMF analysis for sentence (1). The gold standard is compared to the GPM analysis in Fig. 2, and to the JTP analysis in Fig. 3 (we selected one of the two analyses where both parsers don't agree). In addition to the colours that may not be visible in print[8], non-predicted attachments (blue) are additionally labelled as GOLD and falsely predicted attachments (red) are labelled as GPM and JTP respectively. Matches are unmarked.

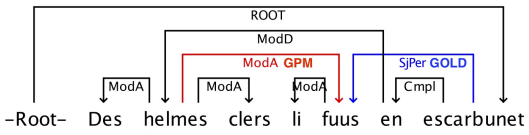


Figure 2: Left dislocation (GOLD vs GPM), cf. (1)

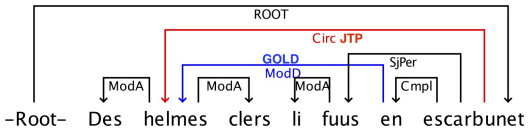


Figure 3: Left dislocation (GOLD vs JTP), cf. (1)

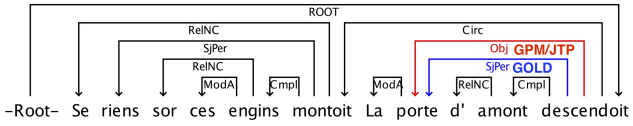
The GPM predicts *ModD* correctly, but misses the subject relation for *li fuus* 'the fire'. It proposes a null-subject structure where *li fuus* modifies *des helmes* (like 'from the helmets of the fire'). The JTP predicts an adjunct (*Circ*) instead of *ModD*, but correctly predicts the other relations.

4.3.2. Subjects and null subjects

Subjects are of two kinds: personal (*SjPer*) and impersonal (*SjImp*). After *ModD*, impersonal subjects exhibit the second highest difference between precision and recall (GPM: 80.0%/45.5%, JTP 87.5%/63.6%). From a linguistic point of view, the difference between both kinds matters: null subjects, impersonal clauses and movement phenomena like stylistic fronting are closely connected. *SjImp* appears only with the impersonal pronoun *il*, which however has the same form as the personal pronoun *il* 'he'. It fills—in null-subject languages optionally—the position of the extraposed structure which appears post-verbally and is labelled *Cmpl* in SRCMF, as in (3).

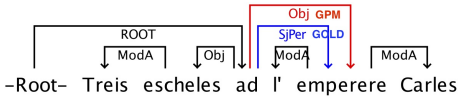
(3) il$_{SjImp}$ m' avint [une grande malaventure$_{Cmpl}$]
it me happened a great misfortune
A great misfortune happened to me.

Apart from the homography of the two pronouns *il*, the second explanation for the low recall is the fact that most of the verbs occurring in impersonal constructions can also be constructed personally. Finally, even for human annotators, it is often hard to draw the line between an impersonal construction like *il est bien drois* 'it is good right' and a copular construction with referential *il* and a subject complement ('this is good right').
With *SjImp*, the JTP clearly outscores the GPM by 7.5 points (precision) and 18.19 points (recall), but the analysis of the examples that were correctly predicted by only one of the parsers has not revealed any generalisable patterns. Contrary to *ModD*, where a number of relations were falsely predicted, *SjPer* was the only alternative to *SjImp*, so that the error was limited to the type of subject. It is easy to calculate the improvement of overall accuracy if this distinction is dropped, but this is not our point here.
Personal subjects (*SjPer*) as such are not a problem: precision and recall are acceptable for both parsers (GPM: 87.43%/88.76%, JTP: 89.40%/90.72%). Rather, the problem is that in OF, as a null-subject language like e.g. Modern Italian or Spanish, the overt realisation of the subject is not a grammatical requirement. Typically, in the false predictions, the subject is mistaken for another category.
This is exemplified by the analysis in Fig. 2 above. By far the most frequent false prediction of both parsers for *SjPer* is the direct object (*Obj*). This is quite understandable with verbs which can have transitive and intransitive constructions, like *descendre* ('sb lowers sth' or 'sth goes down') in Fig. 4, where the intransitive construction, whose only argument is *SjPer*, can be mistaken for a transitive null-subject construction, whose argument is *Obj*.



If something on these constructions climbed the door from up went down

Figure 4: Personal subject (GOLD vs GPM/JTP)



Three divisions has the emperor Charles

Figure 5: Personal subject (GOLD vs GPM)

In some cases, the treatment of coordinations in the CoNLL version of the corpus (see section 3.1.), in combination with null-subject sentences, may cause the issue exemplified in Fig. 5: the parser predicts a direct object even if the verb

[8] Sebastian Riedel's *WhatsWrongWithMyNLP* was used to render the graphs in the figures.

governs two direct objects (*Obj*) in the predicted structure. Structures with two direct objects do not occur in the original version of the SRCMF treebank (a different treatment of the *coord-*... relations could avoid this particular issue, but it creates a number of additional and not very frequent relations which reduce the global LAS by about 0.25 points). The structure in Fig. 5 is predicted by the GPM, whereas the analysis of the JTP is correct.

Although the examples and analyses discussed here have revealed some differences, they have shown that particular relations like dislocation or null-subject sentences pose a problem for both parsers. The hypothesis that the two parsers systematically produce different results with particular syntactic structures was not borne out.

4.4. Parser agreement

The last part of the evaluation is a category-specific calculation of parser agreement. In Table 6, for each category, the number of correct predictions made by *both* parsers features in the second column, followed by the ratio between this number and the total of correct predictions for each parser, GPM in column 3, JTP in column 4. For example, line *ModD* indicates that both parsers correctly predicted three identical instances of *ModD*, i.e. 75% of a total of four correct predictions.

dependency relation	correctly predicted by GPM∩JTP	percentage of total correct predictions GPM	JTP
Apst	89	82.41%	83.18%
AtObj	14	82.35%	63.64%
AtSj	201	77.31%	74.72%
Aux	623	77.68%	77.49%
Circ	1401	53.88%	53.88%
Cmpl	825	71.00%	68.86%
Insrt	37	90.24%	94.87%
Intj	8	88.89%	100.00%
ModA	1675	34.25%	34.26%
ModD	3	75.00%	75.00%
Ng	448	76.06%	75.68%
NgPrt	77	95.06%	96.25%
Obj	1144	65.45%	63.77%
Regim	126	80.77%	87.50%
RelC	418	65.21%	65.01%
RelNC	1673	46.37%	46.28%
Rfc	253	94.05%	92.34%
Rfx	0	0%	0%
SjImp	16	80.00%	57.14%
SjPer	1396	63.08%	62.07%

Table 6: Agreement of correct predictions GPM ∩ JTP

High values of agreement are attained with categories which are bound to a limited number of forms, like negation or interjection. Some of the lower values (60%-70%) concern the arguments which matter most for the verb valency, i.e. subject, direct object (*Obj*), indirect and locative object (*Cmpl*) and adjunct (*Circ*), but the analysis of these cases did not reveal any regularities which would allow me to make plausible linguistic generalisations. The—not very satisfying—conclusion for this experiment is to say that machines seem to have diverging analyses where human annotators have diverging opinions or need more elaborate, often verb-class specific criteria, which are difficult to learn for parsers.

5. Conclusions and Resources

The comparison of two parsers for Old French has confirmed that when chosing a parser for a given language, it is important to consider the syntactic and morphological properties of the language. In the case of Old French, a transition-based joint parser performs better than a graph-based parser. In an in-depth linguistic evaluation of mismatches between the gold standard and the two predictions, it was shown that the transition-based parser outperformed the graph-based parser in some particular categories, but the detailed comparison of the results did not reveal a clear picture from the linguistic point of view.

The parsers used in this experiment are freely available at the sites indicated above. The trained Old French models for both parsers will be made available directly or via links to my website in the LREC repository, and a complete parsing pipeline will be installed on the CLARIN-D platform *WebLicht*.[9]

6. Bibliographical References

Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (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.

Bohnet, B. and Kuhn, J. (2012). The best of bothworlds. a graph-based completion model for transition-based parsers. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–87, Avignon, France, April. Association for Computational Linguistics.

Bohnet, B., Nivre, J., Boguslavsky, I., Farkas, R., Ginter, F., and Hajic, J. (2013). Joint morphological and syntactic analysis for richly inflected languages. In *TACL 1*, pages 415–428.

Bohnet, B. (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.

Guillot, C., Marchello-Nizia, C., and Lavrentiev, A. (2007). La base de français médiéval (bfm): états et perspectives. In Pierre Kunstmann et al., editors, *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*. Steiner, Stuttgart.

Kromann, M. T. and Mikkelsen, L. (2003). The danish dependency treebank and the dtag treebank tool. In Joakim Nivre et al., editors, *Proceedings of the Second Workshop on Treebanks and Linguistic Theories, 14-15 November 2003, Växjö, Sweden*, pages 217–220, Växjö. Växjö University Press.

[9] https://weblicht.sfs.uni-tuebingen.de

Kunstmann, P. and Stein, A. (2007). Le nouveau corpus d'amsterdam. In Pierre Kunstmann et al., editors, *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*, pages 9–27. Steiner, Stuttgart.

Kübler, S., Ryan, M., and Nivre, J. (2009). *Dependency Parsing*. Morgan and Claypool.

Mazziotta, N. (2010). Building the 'syntactic reference corpus of medieval french' using notabene rdf annotation tool. In *Proceedings of the 4th Linguistic Annotation Workshop (LAW IV)*.

Müller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA, October. Association for Computational Linguistics.

Alain Polguère et al., editors. (2009). *Dependency in Linguistic Description*. Benjamins, Amsterdam, Philadelphia.

Schmid, H. (1997). Probabilistic part-of-speech tagging using decision trees. In Daniel Jones et al., editors, *New Methods in Language Processing*, Studies in Computational Linguistics, pages 154–164. UCL Press, London, GB.

Souvay, G. and Pierrel, J.-M. (2010). Lgerm: lemmatization of middle french words lgerm: Lemmatisation des mots en moyen français. *Traitement Automatique des Langues*, 50:149–172.

Stein, A. and Prévost, S. (2013). Syntactic annotation of medieval texts: the syntactic reference corpus of medieval french (SRCMF). In Paul Bennett, et al., editors, *New Methods in Historical Corpora*, Corpus Linguistics and International Perspectives on Language, CLIP Vol. 3, pages 275–282. Narr, Tübingen.

Stein, A. (2007). Resources and tools for old french text corpora. In Yuji Kawaguchi et al., editors, *Corpus-Based Perspectives in Linguistics*, Usage Based Linguistic Informatics; 6, pages 217–229. Benjamins, Amsterdam.

Stein, A. (2014). Parsing heterogeneous corpora with a rich dependency grammar. In Nicoletta Calzolari et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 26.-31.5.2014*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tesnière, L. (1965). *Éléments de syntaxe structurale*. Klincksieck, Paris, 2 edition.

7. Language Resource References

Prévost, Sophie and Stein, Achim. (2013). *Syntactic Reference Corpus of Medieval French (SRCMF)*. ENS de Lyon; Lattice, Paris; Universität Stuttgart, 0.9, ISLRN 899-492-963-833-3.