# Lemmatization and Lexicalized Statistical Parsing of Morphologically Rich Languages: the Case of French

**Djamé Seddah**
Alpage Inria & Univ. Paris-Sorbonne
Paris, France

**Grzegorz Chrupała**
Spoken Language System, Saarland Univ.
Saarbrücken, Germany

**Özlem Çetinoğlu** and **Josef van Genabith**
NCLT & CNGL, Dublin City Univ.
Dublin, Ireland

**Marie Candito**
Alpage Inria & Univ. Paris 7
Paris, France

## Abstract

This paper shows that training a lexicalized parser on a lemmatized morphologically-rich treebank such as the French Treebank slightly improves parsing results. We also show that lemmatizing a similar in size subset of the English Penn Treebank has almost no effect on parsing performance with gold lemmas and leads to a small drop of performance when automatically assigned lemmas and POS tags are used. This highlights two facts: (i) lemmatization helps to reduce lexicon data-sparseness issues for French, (ii) it also makes the parsing process sensitive to correct assignment of POS tags to unknown words.

## 1 Introduction

Large parse-annotated corpora have led to an explosion of interest in statistical parsing methods, including the development of highly successful models for parsing English using the Wall Street Journal Penn Treebank (PTB, (Marcus et al., 1994)). Over the last 10 years, parsing performance on the PTB has hit a performance plateau of 90-92% f-score using the PARSEVAL evaluation metric. When adapted to other language/treebank pairs (such as German, Hebrew, Arabic, Italian or French), to date these models have performed much worse.

A number of arguments have been advanced to explain this performance gap, including limited amounts of training data, differences in treebank annotation schemes, inadequacies of evaluation metrics, linguistic factors such as the degree of word order variation, the amount of morphological information available to the parser as well as the effects of syncretism prevalent in many morphologically rich languages.

Even though none of these arguments in isolation can account for the systematic performance gap, a pattern is beginning to emerge: morphologically rich languages tend to be susceptible to parsing performance degradation.

Except for a residual clitic case system, French does not have explicit case marking, yet its morphology is considerably richer than that of English, and French is therefore a candidate to serve as an instance of a morphologically rich language (MRL) that requires specific treatment to achieve reasonable parsing performance.

Interestingly, French also exhibits a limited amount of word order variation occurring at different syntactic levels including (i) the word level (e.g. pre or post nominal adjective, pre or post verbal adverbs); (ii) phrase level (e.g. possible alternations between post verbal NPs and PPs). In order to avoid discontinuous constituents as well as traces and coindexations, treebanks for this language, such as the French Treebank (FTB, (Abeillé et al., 2003)) or the Modified French Treebank (MFT, (Schluter and van Genabith, 2007)), propose a flat annotation scheme with a non-configurational distinction between adjunct and arguments.

Finally, the extraction of treebank grammars from the French treebanks, which contain less than a third of the annotated data as compared to PTB, is subject to many data sparseness issues that contribute to a performance ceiling, preventing the statistical parsing of French to reach the same level of performance as for PTB-trained parsers (Candito et al., 2009).

This data sparseness bottleneck can be summarized as a problem of optimizing a parsing model along two axes: the grammar and the lexicon. In both cases, the goal is either to get a more compact grammar at the rule level or to obtain a consider-

*Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 85–93,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

ably less sparse lexicon. So far, both approaches have been tested for French using different means and with different degrees of success.

To obtain better grammars, Schluter and van Genabith (2007) extracted a subset of an early release of the FTB and carried out extensive restructuring, extensions and corrections (referred to as the Modified French Treebank MFT) to support grammar acquisition for PCFG-based LFG Parsing (Cahill et al., 2004) while Crabbé and Candito (2008) slightly modified the original FTB POS tagset to optimize the grammar with latent annotations extracted by the Berkeley parser (BKY, (Petrov et al., 2006)).

Moreover, research oriented towards adapting more complex parsing models to French showed that lexicalized models such as Collins' model 2 (Collins, 1999) can be tuned to cope effectively with the flatness of the annotation scheme in the FTB, with the Charniak model (Charniak, 2000) performing particularly well, but outperformed by the BKY parser on French data (Seddah et al., 2009).

Focusing on the lexicon, experiments have been carried out to study the impact of different forms of word clustering on the BKY parser trained on the FTB. Candito et al. (2009) showed that using gold lemmatization provides a significant increase in performance. Obviously, less sparse lexical data which retains critical pieces of information can only help a model to perform better. This was shown in (Candito and Crabbé, 2009) where distributional word clusters were acquired from a 125 million words corpus and combined with inflectional suffixes extracted from the training data. Training the BKY parser with 1000 clusters boosts its performance to the current state-of-the-art with a PARSEVAL $F_1$ score of 88.28% (baseline was 86.29 %).

We performed the same experiment using the CHARNIAK parser and recorded only a small improvement (from 84.96% to 85.51%). Given the fact that lexical information is crucial for lexicalized parsers in the form of bilexical dependencies, this result raises the question whether this kind of clustering is in fact too drastic for lexicalized parsers as it may give rise to head-to-head dependencies which are too coarse. To answer this question, in this paper we explore the impact of lemmatization, as a (rather limited) constrained form of clustering, on a state-of-the-art lexicalized parser (CHARNIAK). In order to evaluate the influence of lemmatization on this parser (which is known to be highly tuned for English) we carry out experiments on both the FTB and on a lemmatized version of the PTB. We used gold lemmatization when available and an automatic statistical morphological analyzer (Chrupała, 2010) to provide more realistic parsing results.

The idea is to verify whether lemmatization will help to reduce data sparseness issues due to the French rich morphology and to see if this process, when applied to English will harm the performance of a parser optimized for the limited morphology of English.

Our results show that the key issue is the way unseen tokens (lemmas or words) are handled by the CHARNIAK parser. Indeed, using pure lemma is equally suboptimal for both languages. On the other hand, feeding the parser with both lemma and part-of-speech slightly enhances parsing performance for French.

We first describe our data sets in Section 2, introduce our data driven morphology process in Section 3, then present experiments in Section 4. We discuss our results in Section 5 and compare them with related research in Section 6 before concluding and outlining further research.

## 2 Corpus

THE FRENCH TREEBANK is the first annotated and manually corrected treebank for French. The data is annotated with labeled constituent trees augmented with morphological annotations and functional annotations of verbal dependents. Its key properties, compared with the PTB, are the following :

*Size:* The FTB consists of 350,931 tokens and 12,351 sentences, that is less than a third of the size of PTB. The average length of a sentence is 28.41 tokens. By contrast, the average sentence length in the Wall Street Journal section of the PTB is 25.4 tokens.

*A Flat Annotation Scheme:* Both the FTB and the PTB are annotated with constituent trees. However, the annotation scheme is flatter in the FTB. For instance, there are no VPs for finite verbs and only one sentential level for clauses or sentences whether or not they are introduced by a complementizer. Only the *verbal nucleus* (VN) is annotated and comprises the verb, its clitics, auxiliaries, adverbs and negation.

*Inflection:* French morphology is richer than English and leads to increased data sparseness for statistical parsing. There are 24,098 lexical types in the FTB, with an average of 16 tokens occurring for each type.

*Compounds:* Compounds are explicitly annotated and very frequent in the treebank: 14.52% of tokens are part of a compound. Following Candito and Crabbé (2009), we use a variation of the treebank where compounds with regular syntactic patterns have been expanded. We refer to this instance as FTB-UC.

*Lemmatization:* Lemmas are included in the treebank's morphological annotations and denote an abstraction over a group of inflected forms. As there is no distinction between semantically ambiguous lexemes at the word form level, polysemic homographs with common inflections are associated with the same lemma (Abeillé et al., 2003). Thus, except for some very rare cases, a pair consisting of a word form and its part-of-speech unambiguously maps to the same lemma.

### 2.1 Lemmatizing the Penn Treebank

Unlike the FTB, the PTB does not have gold lemmas provided within the treebank. We use the finite state morphological analyzer which comes within the English ParGram Grammar (Butt et al., 1999) for lemmatization. For open class words (nouns, verbs, adjectives, adverbs) the word form is sent to the morphological analyzer. The English ParGram morphological analyzer outputs all possible analyses of the word form. The associated gold POS from the PTB is used to disambiguate the result. The same process is applied to closed class words where the word form is different from the lemma (e.g. 'll for will). For the remaining parts of speech the word form is assigned to the lemma.

Since gold lemmas are not available for the PTB, a large-scale automatic evaluation of the lemmatizer is not possible. Instead, we conducted two manual evaluations. First, we randomly extracted 5 samples of 200 <POS,word> pairs from Section 23 of the PTB. Each data set is fed into the lemmatization script, and the output is manually checked. For the 5x200 <POS,word> sets the number of incorrect lemmas is 1, 3, 2, 0, and 2. The variance is small indicating that the results are fairly stable. For the second evaluation, we extracted each unseen word from Section 23 and manually checked the accuracy of the lemmatization. Of the total of 1802 unseen words, 394 words are associated with an incorrect lemma (331 unique) and only 8 with an incorrect <POS,lemma> pair (5 unique). For an overall unseen word percentage of 3.22%, the lemma accuracy is 77.70%. If we assume that all seen words are correctly lemmatized, overall accuracy would be 99.28%.

### 2.2 Treebank properties

In order to evaluate the influence of lemmatization on comparable corpora, we extracted a random subset of the PTB with properties comparable to the FTB-UC (mainly with respect to CFG size and number of tokens). We call this PTB subset S.PTB. Table 1 presents a summary of some relevant features of those treebanks.

| | FTBUC | S.PTB | PTB |
|---|---:|---:|---:|
| # of tokens | 350,931 | 350,992 | 1,152,305 |
| # of sentences | 12,351 | 13,811 | 45,293 |
| average length | 28,41 | 25.41 | 25.44 |
| CFG size | 607,162 | 638,955 | 2,097,757 |
| # unique CFG rules | 43,413 | 46,783 | 91,027 |
| # unique word forms | 27,130 | 26,536 | 47,678 |
| # unique lemmas | 17,570 | 20,226 | 36,316 |
| ratio words/lemma | 1.544 | 1.311 | 1.312 |

Table 1: French and Penn Treebanks properties

Table 1 shows that the average number of word forms associated with a lemma (i.e. the lemma ratio) is higher in the FTB-UC (1.54 words/lemma) than in the PTB (1.31). Even though the PTB ratio is lower, it is still large enough to suggest that even the limited English morphology should be taken into account when aiming at reducing lexicon sparseness.

Trying to learn French and English morphology in a data driven fashion in order to predict lemma from word forms is the subject of the next section.

## 3 Morphology learning

In order to assign morphological tags and lemmas to words we use the MORFETTE model (Chrupała, 2010), which is a variation of the approach described in (Chrupała et al., 2008).

MORFETTE is a sequence labeling model which combines the predictions of two classification models (one for morphological tagging and one for lemmatization) at decoding time, using beam search.

### 3.1 Overview of the Morfette model

The morphological classes correspond simply to the (fine-grained) POS tags. Lemma classes are edit scripts computed from training data: they specify which string manipulations (such as character deletions and insertions) need to be performed in order to transform the input string (word form) into the corresponding output string (lemma).

The best sequence of lemmas and morphological tags for input sentence $\mathbf{x}$ is defined as:

$$(\hat{\mathbf{l}}, \hat{\mathbf{m}}) = \arg\max_{(\mathbf{l},\mathbf{m})} P(\mathbf{l}, \mathbf{m}|\mathbf{x})$$

The joint probability is decomposed as follows:

$$\begin{aligned} P(l_0 ... l_i, m_0 ... m_i|\mathbf{x}) =& P_L(l_i|m_i, \mathbf{x}) P_M(m_i|\mathbf{x}) \\ & \times P(m_0 ... m_{i-1}, l_0 ... l_{i-1}|\mathbf{x}) \end{aligned}$$

where $P_L(l_i|m_i, \mathbf{x})$ is the probability of lemma class $l$ at position $i$ according to the lemma classifier, $P_M(m_i|\mathbf{x})$ is the probability of the tag $m$ at position $i$ according to the morphological tag classifier, and $\mathbf{x}$ is the sequence of words to label.

While Chrupała et al. (2008) use Maximum Entropy training to learn $P_M$ and $P_L$, here we learn them using Averaged Perceptron algorithm due to Freund and Schapire (1999). It is a much simpler algorithm which in many scenarios (including ours) performs as well as or better than MaxEnt.

We also use the general Edit Tree instantiation of the edit script as developed in (Chrupała, 2008). We find the longest common substring (LCS) between the form $w$ and the lemma $w'$. The portions of the string in the word form before (prefix) and after (suffix) the LCS need to be modified in some way, while the LCS (stem) stays the same. If there is no LCS, then we simply record that we need to replace $w$ with $w'$ . As for the modifications to the prefix and the suffix, we apply the same procedure recursively: we try to find the LCS between the prefix of $w$ and the prefix of $w'$. If we find one, we recurse; if we do not, we record the replacement; we do the same for the suffix.

### 3.2 Data Set

We trained MORFETTE on the standard splits of the FTB with the first 10% as test set, the next 10% for the development set and the remaining for training (i.e. 1235/1235/9881 sentences). Lemmas and part-of-speech tags are given by the treebank annotation scheme.

As pointed out in section 2.1, PTB's lemmas have been automatically generated by a deterministic process, and only a random subset of them have been manually checked. For the remainder of this paper, we treat them as gold, regardless of the errors induced by our PTB lemmatizer.

The S.PTB follows the same split as the FTB-UC, first 10% for test, next 10% for dev and the last 80% for training (i.e. 1380/1381/11050 sentences).

MORFETTE can optionally use a morphological lexicon to extract features. For French, we used the extended version of Le*fff* (Sagot et al., 2006) and for English, the lexicon used in the Penn XTAG project (Doran et al., 1994). We reduced the granularity of the XTAG tag set, keeping only the bare categories. Both lexicons contain around 225 thousands word form entries.

### 3.3 Performance on French and English

Table 2 presents results of MORFETTE applied to the development and test sets of our treebanks. Part-of-speech tagging performance for French is state-of-the-art on the FTB-UC, with an accuracy of 97.68%, on the FTB-UC test set, only 0.02 points (absolute) below the MaxEnt POS tagger of Denis and Sagot (2009). Comparing MORFETTE's tagging performance for English is a bit more challenging as we only trained on one third of the full PTB and evaluated on approximately one section, whereas results reported in the literature are usually based on training on sections 02-18 and evaluating on either sections 19-21 or 22-24. For this setting, state-of-the-art POS accuracy for PTB tagging is around 97.33%. On our PTB sample, MORFETTE achieves 96.36% for all words and 89.64 for unseen words.

Comparing the lemmatization performance for both languages on the same kind of data is even more difficult as we are not aware of any data driven lemmatizer on the same data. However, with an overall accuracy above 98% for the FTB-UC (91.5% for unseen words) and above 99% for the S.Ptb (95% for unseen words), lemmatization performs well enough to properly evaluate parsing on lemmatized data.

| | FTBUC | | S.PTB | |
|---|---|---|---|---|
| Dev | All | Unk. (4.8) | All | Unk. (4.67) |
| POS acc | 97.38 | 91.95 | 96.36 | 88.90 |
| Lemma acc | 98.20 | 92.52 | 99.11 | 95.51 |
| Joint acc | 96.35 | 87.16 | 96.26 | 87.05 |
| Test | All | Unk. (4.62) | All | Unk. (5.04) |
| POS acc | 97.68 | 90.52 | 96.53 | 89.64 |
| Lemma acc | 98.36 | 91.54 | 99.13 | 95.72 |
| Joint acc | 96.74 | 85.28 | 96.45 | 88.49 |

Table 2: POS tagging and lemmatization performance on the Ftb and on the S.Ptb

## 4 Parsing Experiments

In this section, we present the results of two sets of experiments to evaluate the impact of lemmatization on the lexicalized statistical parsing of two languages, one morphologically rich (French), but with none of its morphological features exploited by the Charniak parser, the other (English) being quite the opposite, with the parser developed mainly for this language and Ptb annotated data. We show that lemmatization results in increased performance for French, while doing the same for English penalizes parser performance.

### 4.1 Experimental Protocol

**Data** The data sets described in section 3.2 are used throughout. The version of the Charniak parser (Charniak, 2000) was released in August 2005 and recently adapted to French (Seddah et al., 2009).

**Metrics** We report results on sentences of length less than 40 words, with three evaluation metrics: the classical ParsEval Labeled brackets $F_1$ score, POS tagging accuracy (excluding punctuation tags) and the Leaf Ancestor metric (Sampson and Babarczy, 2003) which is believed to be somewhat more neutral with respect to the treebank annotation scheme than ParsEval (Rehbein and van Genabith, 2007).

**Treebank tag sets** Our experiments involve the inclusion of POS tags directly in tokens. We briefly describe our treebank tag sets below.

- Ftb-uc tag set: "cc" This is the tag set developed by (Crabbé and Candito, 2008) (Table 4), known to provide the best parsing performance for French (Seddah et al., 2009). Like in the Ftb, preterminals are the main categories, but they are also augmented with a WH flag for A, ADV, PRO and with the mood for verbs (there are 6 moods). No information is propagated to non-terminal symbols.

ADJ ADJWH ADV ADVWH CC CLO CLR CLS CS DET
DETWH ET I NC NPP P P+D P+PRO PONCT PREF PRO
PROREL PROWH V VIMP VINF VPP VPR VS

Table 4: cc tag set

- The Ptb tag set This tag set is described at length in (Marcus et al., 1994) and contains supplementary morphological information (e.g. number) over and above what is represented in the cc tag set for French. Note that some information is marked at the morphological level in English (superlative, "the greatest (JJS)") and not in French (" le plus (ADV) grand (ADJ)").

CC CD DT EX FW IN JJ JJR JJS LS MD NN NNP NNPS
NNS PDT POS PRP PRP$ RB RBR RBS RP SYM TO UH
VB VBD VBG VBN VBP VBZ WDT WP WP$ WRB

Table 5: Ptb tag set

### 4.2 Cross token variation and parsing impact

From the source treebanks, we produce 5 versions of tokens: tokens are generated as either simple POS tag, gold lemma, gold lemma+gold POS, word form, and word form+gold POS. The token versions successively add more morphological information. Parsing results are presented in Table 3.

**Varying the token form** The results show that having no lexical information at all (POS-only) results in a small drop of ParsEval performance for French compared to parsing lemmas, while the corresponding Leaf Ancestor score is actually higher. For English having no lexical information at all leads to a drop of 2 points in ParsEval. The *so-called* impoverished morphology of English appears to bring enough morphological information to raise tagging performance to 95.92% (from POS-only to word-only).

For French the corresponding gain is only 2 points of POS tagging accuracy. Moreover, between these

| | French Treebank UC | | | Sampled Penn Treebank | | |
|---|---|---|---|---|---|---|
| Tokens | $F_1$ score | Pos acc. | leaf-Anc. | $F_1$ score | Pos acc. | leaf-Anc. |
| POS-only | 84.48 | 100 | 93.97 | 85.62 | 100 | 94.02 |
| lemma-only | 84.77 | 94.23 | 93.76 | 87.69 | 89.22 | 94.92 |
| word-only | 84.96 | 96.26 | 94.08 | 88.64 | 95.92 | 95.10 |
| **lemma-POS** | **86.83**$^{(1)}$ | **98.79** | **94.65** | **89.59**$^{(3)}$ | **99.97** | **95.41** |
| word-POS | 86.13$^{(2)}$ | 98.4 | 94.46 | 89.53$^{(4)}$ | 99.96 | 95.38 |

Table 3: Parsing performance on the FTB-UC and the S.PTB with tokens variations using gold lemmas and gold POS. *( p-value* $(1)$ *&* $(2) = 0.007$*; p-value* $(3)$ *&* $(4) = 0.146$*. All other configurations are statistically significant.)*

two tokens variations, POS-only and word-only, parsing results gain only half a point in PARSEVAL and almost nothing in leaf Ancestor.

Thus, it seems that encoding more morphology (i.e. including word forms) in the tokens does not lead to much improvement for parsing French as opposed to English. The reduction in data sparseness due to the use of lemmas alone is thus not sufficient to counterbalance the lack of morphological information.

However, the large gap between POS tagging accuracy seen between lemma-only and word-only for English indicates that the parser makes use of this information to provide at least reasonable POS guesses.

For French, only 0.2 points are gained for PARSEVAL results between lemma-only to word-only, while POS accuracy benefits a bit more from including richer morphological information.

This raises the question whether the FTB-UC provides enough data to make its richer morphology informative enough for a parsing model.

**Suffixing tokens with POS tags** It is only when gold POS are added to the lemmas that one can see the advantage of a reduced lexicon for French. Indeed, performance peaks for this setting (lemma-POS). The situation is not as clear for English, where performance is almost identical when gold POS are added to lemmas or words. POS Tagging is nearly perfect, thus a performance ceiling is reached. The very small differences between those two configurations (most noticeable with the Leaf Ancestor score of 95.41 vs. 95.38) indicates that the reduced lemma lexicon is actually of some limited use but its impact is negligible compared to perfect tagging.

While the lemma+POS setting clearly boosts performance for parsing the FTB, the situation is less clear for English. Indeed, the lemma+POS and the word+POS gold variations give almost the same results. The fact that the POS tagging accuracy is close to 100% in this mode shows that the key parameter for optimum parsing performance in this experiment is the ability to guess POS for unknown words well.

In fact, the CHARNIAK parser uses a two letter suffix context for its tagging model, and when gold POS are suffixed to any type of token (being lemma or word form), the PTB POS tagset is used as a substitute for lack of morphology.

It should also be noted that the FTB-UC tag set does include some discriminative features (such as PART, INF and so on) but those are expressed by more than two letters, and therefore a two letter suffix tag cannot really be useful to discriminate a richer morphology. For example, in the PTB, the suffix BZ, as in VBZ, always refers to a verb, whereas the FTB pos tag suffix PP, as in NPP (Proper Noun) is also found in POS labels such as VPP (past participle verb).

### 4.3 Realistic Setup: Using Morfette to help parsing

Having shown that parsing French benefits from a reduced lexicon is not enough as results imply that a key factor is POS tag guessing. We therefore test our hypothesis in a more realistic set up. We use MORFETTE to lemmatize and tag raw words (instead of the "gold" lemma-based approach described above), and the resulting corpus is then parsed using the corresponding training set.

In order to be consistent with PARSEVAL POS evaluation, which does not take punctuation POS into account, we provide a summary of MORFETTE's performance for such a configuration in (Table 6).

Results shown in Table 7 confirm our initial hy-

| | POS acc | Lemma acc | Joint acc |
|---|---|---|---|
| FTB-UC | 97.34 | 98.12 | 96.26 |
| S.PTB | 96.15 | 99.04 | 96.07 |

Table 6: PARSEVAL Pos tagging accuracy of treebanks test set

pothesis for French. Indeed, parsing performance peaks with a setup involving automatically generated lemma and POS pairs, even though the difference with raw words+auto POS is not statistically significant for the PARSEVAL $F_1$ metric[1]. Note that parser POS accuracy does not follow this pattern. It is unclear exactly why this is the case. We speculate that the parser is helped by the reduced lexicon but that performance suffers when a <lemma,POS> pair has been incorrectly assigned by MORFETTE, leading to an increase in unseen tokens. This is confirmed by parsing the same lemma but with gold POS. In that case, parsing performance does not suffer too much from CHARNIAK's POS guessing on unseen data.

For the S.PTB, results clearly show that both the automatic <lemma,POS> and <word,POS> configurations lead to very similar results (yet statistically significant with a $F_1$ $p$-value = 0.027); having the same POS accuracy indicates that most of the work is done at the level of POS guessing for unseen tokens, and in this respect the CHARNIAK parser clearly takes advantage of the information included in the PTB tag set.

| | $F_1$ score | Pos acc. | leaf-Anc. |
|---|---|---|---|
| S.PTB | | | |
| auto lemma only | 87.11 | 89.82 | 94.71 |
| auto lemma+auto pos (a) | 88.15 | 96.21 | 94.85 |
| **word +auto pos** (b) | **88.28** | **96.21** | **94.88** |
| *$F_1$ p-value: (a) and (b)* | *0.027* | | |
| *auto lemma+gold pos* | *89.51* | *99.96* | *95,36* |
| FTB-UC | | | |
| auto lemma only | 83.92 | 92.98 | 93.53 |
| **auto lemma+auto pos** (c) | **85.06** | 96.04 | **94.14** |
| word +auto pos (d) | 84.99 | **96.47** | 94.09 |
| *$F_1$ p-value: (c) and (d)* | *0.247* | | |
| *auto lemma+gold pos* | *86.39* | *97.35* | *94.68* |

Table 7: Realistic evaluation of parsing performance

## 5 Discussion

When we started this work, we wanted to explore the benefit of lemmatization as a means to reduce data sparseness issues underlying statistical lexicalized parsing of small treebanks for morphologically rich languages, such as the FTB. We showed that the expected benefit of lemmatization, a less sparse lexicon, was in fact hidden by the absence of inflectional information, as required by e.g. the CHARNIAK parser to provide good POS guesses for unseen words. Even the inclusion of POS tags generated by a state-of-the-art tagger (MORFETTE) did not lead to much improvement compared to a parser run in a regular bare word set up.

An unexpected effect is that the POS accuracy of the parser trained on the French data does not reach the same level of performance as our tagger (96.47% for <word, auto POS> vs. 97.34% for MORFETTE). Of course, extending the CHARNIAK tagging model to cope with lemmatized input should be enough, because its POS guessing model builds on features such as capitalization, hyphenation and a two-letter suffix (Charniak, 2000). Those features are not present in our current lemmatized input and thus cannot be properly estimated.

CHARNIAK also uses the probability that a given POS is realized by a previously unobserved word. If any part of a <lemma,POS> pair is incorrect, the number of unseen words in the test set would be higher than the one estimated from the training set, which only contained correct lemmas and POS tags in our setting. This would lead to unsatisfying POS accuracy. This inadequate behavior of the unknown word tagging model may be responsible for the POS accuracy result for <auto lemma> (cf. Table 7, lines <auto lemma only> for both treebanks).

We believe that this performance degradation (or in this case the somewhat less than expected improvement in parsing results) calls for the inclusion of all available lexical information in the parsing model. For example, nothing prevents a parsing model to condition the generation of a head upon a lemma, while the probability to generate a POS would depend on both morphological features and (potentially) the supplied POS.

[1] Statistical significance is computed using Dan Bikel's stratified shuffling implementation: `www.cis.upenn.edu/~dbikel/software.html`.

## 6 Related Work

A fair amount of recent research in parsing morphologically rich languages has focused on coping with unknowns words and more generally with the small and limited lexicons acquired from treebanks. For instance, Goldberg et al. (2009) augment the lexicon for a generative parsing model by including lexical probabilities coming from an external lexicon. These are estimated using an HMM tagger with Baum-Welch training. This method leads to a significant increase of parsing performance over previously reported results for Modern Hebrew. Our method is more stratified: external lexical resources are included as features for Morfette and therefore are not directly seen by the parser besides generated lemma and POS.

For parsing German, Versley and Rehbein (2009) cluster words according to linear context features. The clusters are then integrated as features to boost a discriminative parsing model to cope with unknown words. Interestingly, they also include all possible information: valence information, extracted from a lexicon, is added to verbs and preterminal nodes are annotated with case/number. This leads their discriminative model to state-of-the-art results for parsing German.

Concerning French, Candito and Crabbé (2009) present the results of different clustering methods applied to the parsing of Ftb with the Bky parser. They applied an unsupervised clustering algorithm on the 125 millions words “Est Republicain” corpus to get a reduced lexicon of 1000 clusters which they then augmented with various features such as capitalization and suffixes. Their method is the best current approach for the probabilistic parsing of French with a $F_1$ score (<=40) of 88.29% on the standard test set. We run the Charniak parser on their clusterized corpus. Table 8 summarizes the current state-of-the-art for lexicalized parsing on the Ftb-uc.[2]
Clearly, the approach consisting in extending clusters with features and suffixes seems to improve Charniak’s performance more than our method. In that case, the lexicon is drastically reduced, as well as the amount of out of vocabulary words (OOVs). Nevertheless, the relatively low POS accuracy, with only 36 OOVs, for this configuration confirms that POS guessing is the current bottleneck if a process of reducing the lexicon increases POS assignment ambiguities.

| tokens | $F_1$ | Pos acc | % of OOVs |
|---:|---|---|---|
| raw word (a) | 84.96 | 96.26 | 4.89 |
| auto <lemma,pos> (b) | 85.06 | 96.04 | 6.47 |
| disinflected (c) | 85.45 | 96.51 | 3.59 |
| cluster+caps+suffixes (d) | 85.51 | 96.89 | 0.10 |

Table 8: Charniak parser performance summary on the Ftb-uc test set *(36340 tokens). Compared to (a), all $F_1$ results, but (b), are statistically significant ($p$-values < 0.05), differences between (c) & (d), (b) & (c) and (b) & (d) are not ($p$-values are resp. 0.12, 0.41 and 0.11). Note that the (b) & (d) $p$-value for all sentences is of 0.034, correlating thus the observed gap in parsing performance between these two configuration.*

## 7 Conclusion

We showed that while lemmatization can be of some benefit to reduce lexicon size and remedy data sparseness for a MRL such as French, the key factor that drives parsing performance for the Charniak parser is the amount of unseen words resulting from the generation of <lemma,POS> pairs for the Ftb-uc. For a sample of the English PTB, morphological analysis did not produce any significant improvement.
Finally, even if this architecture has the potential to help out-of-domain parsing, adding morphological analysis on top of an existing highly tuned statistical parsing system can result in suboptimal performance. Thus, in future we will investigate tighter integration of the morphological features with the parsing model.

## Acknowledgments

D. Seddah and M. Candito were supported by the ANR Sequoia (ANR-08-EMER-013); Ö. Çetinoğlu and J. van Genabith by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at Dublin City University; G. Chrupała by BMBF project NL-Search (contract 01IS08020B).

[2] For this comparison, we also trained the Charniak parser on a *disinflected* variation of the Ftb-uc. *Disinflection* is a deterministic, lexicon based process, standing between stemming and lemmatization, which preserves POS assignment ambiguities (Candito and Crabbé, 2009).

## References

Anne Abeillé, Lionel Clément, and François Toussenel, 2003. *Building a Treebank for French*. Kluwer, Dordrecht.

Miriam Butt, María-Eugenia Niño, and Frédérique Segond. 1999. *A Grammar Writer's Cookbook*. CSLI Publications, Stanford, CA.

Aoife Cahill, Michael Burke, Ruth O'Donovan, Josef van Genabith, and Andy Way. 2004. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 320–327, Barcelona, Spain.

Marie Candito and Benoît Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 138–141, Paris, France, October. Association for Computational Linguistics.

Marie Candito, Benoit Crabbé, and Djamé Seddah. 2009. On statistical parsing of french with supervised and semi-supervised strategies. In *EACL 2009 Workshop Grammatical inference for Computational Linguistics*, Athens, Greece.

Eugene Charniak. 2000. A maximum entropy inspired parser. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, pages 132–139, Seattle, WA.

Grzegorz Chrupała, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with morfette. In *In Proceedings of LREC 2008*, Marrakech, Morocco. ELDA/ELRA.

Grzegorz Chrupała. 2008. *Towards a machine-learning architecture for lexical functional grammar parsing*. Ph.D. thesis, Dublin City University.

Grzegorz Chrupała. 2010. Morfette: A tool for supervised learning of morphology. `http://sites.google.com/site/morfetteweb/`. Version 0.3.1.

Michael Collins. 1999. *Head Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Benoit Crabbé and Marie Candito. 2008. Expériences d'analyse syntaxique statistique du français. In *Actes de la 15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'08)*, pages 45–54, Avignon, France.

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proc. of PACLIC*, Hong Kong, China.

Christy Doran, Dania Egedi, Beth Ann Hockey, B. Srinivas, and Martin Zaidel. 1994. Xtag system: A wide coverage grammar for english. In *Proceedings of the 15th conference on Computational linguistics*, pages 922–928, Morristown, NJ, USA. Association for Computational Linguistics.

Yoav Freund and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296.

Yoav Goldberg, Reut Tsarfaty, Meni Adler, and Michael Elhadad. 2009. Enhancing unlexicalized parsing performance using a wide coverage lexicon, fuzzy tag-set mapping, and EM-HMM-based lexical probabilities. In *Proc. of EACL-09*, pages 327–335, Athens, Greece.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn TreeBank. *Computational Linguistics*, 19(2):313–330.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, July. Association for Computational Linguistics.

Ines Rehbein and Josef van Genabith. 2007. Treebank annotation schemes and parser evaluation for german. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague.

Benoit Sagot, Lionel Clément, Eric V. de La Clergerie, and Pierre Boullier. 2006. The lefff 2 syntactic lexicon for french: Architecture, acquisition, use. *Proc. of LREC 06, Genoa, Italy*.

Geoffrey Sampson and Anna Babarczy. 2003. A test of the leaf-ancestor metric for parse accuracy. *Natural Language Engineering*, 9(04):365–380.

Natalie Schluter and Josef van Genabith. 2007. Preparing, restructuring, and augmenting a French Treebank: Lexicalised parsers or coherent treebanks? In *Proc. of PACLING 07*, Melbourne, Australia.

Djamé Seddah, Marie Candito, and Benoit Crabbé. 2009. Cross parser evaluation and tagset variation: A French Treebank study. In *Proceedings of the 11th Internation Conference on Parsing Technologies (IWPT'09)*, pages 150–161, Paris, France, October. Association for Computational Linguistics.

Yannick Versley and Ines Rehbein. 2009. Scalable discriminative parsing for german. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 134–137, Paris, France, October. Association for Computational Linguistics.