Statistical Parsing of Spanish and Data Driven Lemmatization

Joseph Le Roux[†] **Benoît Sagot**[⋆] **Djamé Seddah**[⋆,◇]
[†] Laboratoire d'Informatique Paris Nord, Université Paris Nord, UMR CNRS 7030
[⋆]Alpage, INRIA & Université Paris Diderot
[◇] Université Paris Sorbonne
leroux@univ-paris13.fr, benoit.sagot@inria.fr, djame.seddah@paris-sorbonne.fr

Abstract

Although parsing performances have greatly improved in the last years, grammar inference from treebanks for morphologically rich languages, especially from small treebanks, is still a challenging task. In this paper we investigate how state-of-the-art parsing performances can be achieved on Spanish, a language with a rich verbal morphology, with a non-lexicalized parser trained on a treebank containing only around 2,800 trees. We rely on accurate part-of-speech tagging and data-driven lemmatization to provide parsing models able to cope lexical data sparseness. Providing state-of-the-art results on Spanish, our methodology is applicable to other languages with high level of inflection.

1 Introduction

Grammar inference from treebanks has become the standard way to acquire rules and weights for parsing devices. Although tremendous progress has been achieved in this domain, exploiting small treebanks is still a challenging task, especially for languages with a rich morphology. The main difficulty is to make good generalizations from small example sets exhibiting data sparseness. This difficulty is even greater when the inference process relies on semi-supervised or unsupervised learning techniques which are known to require more training examples, as these examples do not explicitly contain all the information.

In this paper we want to explore how we can cope with this difficulty and get state-of-the-art syntactic analyses with a non-lexicalized parser that uses modern semisupervised inference techniques. We rely on accurate data-driven lemmatization and part-of-speech tagging to reduce data sparseness and ease the burden on the parser. We try to see how we can improve parsing structure predictions solely by modifying the terminals and/or the preterminals of the trees. We keep the rest of the tagset as is.
In order to validate our method, we perform experiments on the Cast3LB constituent treebank for Spanish (Castillan). This corpus is quite small, around 3,500 trees, and Spanish is known to have a rich verbal morphology, making the tag set quite complex and difficult to predict. Cowan and Collins (2005) and Chrupała (2008) already showed interesting results on this corpus that will provide us with a comparison for this work, especially on the lexical aspects as they used lexicalized frameworks while we choose PCFG-LAs.

This paper is structured as follows. In Section 2 we describe the Cast3LB corpus in details. In Section 3 we present our experimental setup and results which we discuss and compare in Section 4. Finally, Section 5 concludes the presentation.

2 Data Set

The Castillan 3LB treebank (Civit and Martì, 2004) contains 3,509 constituent trees with functional annotations. It is divided in training (2,806 trees), development (365 trees) and test (338 trees).

We applied the transformations of Chrupała (2008) to the corpus where CP and SBAR nodes are added to the subordinate and relative clauses but we did not perform any other transformations, like the coordination modification applied by Cowan and Collins (2005).

The Cast3LB tag set is rich. In particular part-of-speech (POS) tags are fine-grained and encode precise morphological information while non-terminal tags describe subcategorization and function labels.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 55–61,
Jeju, Republic of Korea, 12 July 2012.

Without taking functions into account, there are 43 non-terminal tags. The total tag set thus comprises 149 symbols which makes the labeling task challenging.

The rich morphology of Spanish can be observed in the treebank through word form variation. Table 1 shows some figures extracted from the corpus (training, development and test). In particular the word form/lemma ratio is 1.54, which is similar to other Romance language treebanks (French FTB and Italian ITB).

# of tokens	94 907
# of unique word forms	17 979
# of unique lemmas	11 642
ratio word form/lemma	1.54

Table 1: C3LB properties

Thus, we are confronted with a small treebank with a rich tagset and a high word diversity. All these conditions make the corpus a case in point for building a parsing architecture for morphologically-rich languages.

3 Experiments

We conducted experiments on the Cast3LB development set in order to test various treebank modifications, that can be divided in two categories: (i) modification of the preterminal symbols of the treebank by using simplified POS tagsets; (ii) modification of the terminal symbols of the treebank by replacing word tokens by lemmas.

3.1 Experimental Setup

In this section we describe the parsing formalism and POS tagging settings used in our experiments.

PCFG-LAs To test our hypothesis, we use the grammatical formalism of Probabilistic Context-Free Grammars with Latent Annotations (PCFG-LAs) (Matsuzaki et al., 2005; Petrov et al., 2006). These grammars depart from the standard PCFGs by automatically refining grammatical symbols during the training phase, using unsupervised techniques. They have been applied successfully to a wide range of languages, among which French (Candito and Seddah, 2010), German (Petrov and Klein, 2008), Chinese and Italian (Lavelli and Corazza, 2009).

For our experiments, we used the LORG PCFG-LA parser implementing the CKY algorithm. This software also implements the techniques from Attia et al. (2010) for handling out-of-vocabulary words, where interesting suffixes for part-of-speech tagging are collected on the training set, ranked according to their information gain with regards to the part-of-speech tagging task. Hence, all the experiments are presented in two settings. In the first one, called *generic*, unknown words are replaced with a dummy token `UNK`, while in the second one, dubbed *IG*, we use the collected suffixes and typographical information to type unknown words.[1] We retained the 30 best suffixes of length 1, 2 and 3.

The grammar was trained using the algorithm of Petrov and Klein (2007) using 3 rounds of split/merge/smooth[2]. For lexical rules, we applied the strategy dubbed *simple lexicon* in the Berkeley parser. Rare words – words occurring less than 3 times in the training set – are replaced by a special token, which depends on the OOV handling method (*generic* or *IG*), before collecting counts.

POS tagging We performed parsing experiments with three different settings regarding POS information provided as an input to the parser: (i) with no POS information, which constitutes our baseline; (ii) with gold POS information, which can be considered as a topline for a given parser setting; (iii) with POS information predicted using the MElt POS-tagger (Denis and Sagot, 2009), using three different tagsets that we describe below.

MElt is a state-of-the-art sequence labeller that is trained on both an annotated corpus and an external lexicon. The standard version of MElt relies on Maximum-Entropy Markov models (MEMMs). However, in this work, we have used a multiclass perceptron instead, as it allows for much faster training with very small performance drops (see Table 2). For training purposes, we used the training section of the Cast3LB (76,931 tokens) and the Le*ff*e lexicon (Molinero et al., 2009), which contains almost 800,000 distinct (form, category) pairs.[3]

We performed experiments using three different

[1] Names *generic* and *IG* originally come from Attia et al. (2010).

[2] We tried to perform 4 and 5 rounds but 3 rounds proved to be optimal on this corpus.

[3] Note that MElt does not use information from the exter-

TAGSET	baseline	reduced2	reduced3
Nb. of tags	106	42	57
Multiclass Perceptron			
Overall Acc.	96.34	97.42	97.25
Unk. words Acc.	91.17	93.35	92.30
Maximum-Entropy Markov model (MEMM)			
Overall Acc.	96.46	97.42	97.25
Unk. words Acc.	91.57	93.76	92.87

Table 2: MElt POS tagging accuracy on the Cast3LB development set for each of the three tagsets. We provide results obtained with the standard MElt algorithm (MEMM) as well as with the multiclass perceptron, used in this paper, for which training is two orders of magnitude faster. Unknown words represent as high as 13.5 % of all words.

tagsets: (i) a *baseline tagset* which is identical to the tagset used by Cowan and Collins (2005) and Chrupała (2008); with this tagset, the training corpus contains 106 distinct tags;
(ii) the *reduced2* tagset, which is a simplification of the baseline tagset: we only retain the first two characters of each tag from the baseline tagset; with this tagset, the training corpus contains 42 distinct tags;
(iii) the *reduced3* tagset, which is a variant of the reduced2 tagset: contrarily to the reduced2 tagset, the reduced3 tagset has retained the mood information for verb forms, as it proved relevant for improving parsing performances as shown by (Cowan and Collins, 2005); with this tagset, the training corpus contains 57 distinct tags.

Melt POS tagging accuracy on the Cast3LB development set for these three tagsets is given in Table 2, with overall figures together with figures computed solely on unknown words (words not attested in the training corpus, i.e., as high as 13.5 % of all tokens).

3.2 Baseline

The first set of experiments was conducted with the baseline POS tagset. Results are summarized in Table 3. This table presents parsing statistics on the Cast3LB development set in the 3 POS settings introduced above (i) no POS provided, (ii) gold POS provided and (iii) predicted POS provided. For each POS tagging setting it shows labeled precision, labeled recall, labeled F1-score, the percentage of exact match and the POS tagging accuracy. The latter needs not be the same as presented in Section 3.1 because (i) punctuation is ignored and (ii) if the parser cannot use the information provided by the tagger, it is discarded and the parser performs POS-tagging on its own.

MODEL	LP	LR	F1	EXACT	POS
Word Only					
Generic	81.42	81.04	**81.23**	**14.47**	**90.89**
IG	80.15	79.60	79.87	14.19	85.01
Gold POS					
Generic	87.83	87.49	**87.66**	**30.59**	99.98
IG	86.78	86.53	86.65	27.96	99.98
Pred. POS					
Generic	84.47	84.39	**84.43**	**22.44**	95.82
IG	83.60	83.66	83.63	21.78	95.82

Table 3: Baseline PARSEVAL scores on Cast3LB dev. set ($\leq$ 40 words)

As already mentioned above, this tagset contains 106 distinct tags. On the one hand it means that POS tags contain useful information. On the other hand it also means that the data is already sparse and adding more sparseness with the IG suffixes and typographical information is detrimental. This is a major difference between this POS tagset and the two following ones.

3.3 Using simplified tagsets

We now turn to the modified tagsets and measure their impact on the quality of the syntactic analyses. Results are summarized in Table 4 for the *reduced2* tagset and in Table 5 for *reduced3*. In these two settings, we can make the following remarks.

- Parsing results are better with *reduced3*, which indicates that verbal mood is an important feature for correctly categorizing verbs at the syntactic level.
- When POS tags are not provided, using suffixes and typographical information improves OOV word categorization and leads to a better tagging accuracy and F1 parsing score (78.94 vs. 81.81 for *reduced2* and 79.69 vs. 82.44 for *reduced3*).

nal lexicon as constraints, but as features. Therefore, the set of categories in the external lexicon need not be identical to the tagset. In this work, the L*effe* categories we used include some morphological information (84 distinct categories).

- When providing the parser with POS tags, whether gold or predicted, both settings show an interesting difference w.r.t. to unknown words handling. When using *reduced2*, the IG setting is better than the generic one, whereas the situation is reversed in *reduced3*. This indicates that *reduced2* is too coarse to help finely categorizing unknown words and that the refinement brought by IG is beneficial, however the added sparseness. For *reduced3* it is difficult to say whether it is the added richness of the POS tagset or the induced OOV sparseness that explains why IG is detrimental.

MODEL	LP	LR	F1	EXACT	POS
Word Only					
Generic	78.86	79.02	78.94	15.23	88.18
IG	81.89	81.72	**81.81**	**16.17**	**92.19**
Gold POS					
Generic	86.56	85.90	86.23	26.64	100.00
IG	86.90	86.63	**86.77**	**29.28**	100.00
Pred. POS					
Generic	84.16	83.81	83.99	21.05	96.76
IG	84.57	84.32	**84.45**	**21.38**	96.76

Table 4: PARSEVAL scores on Cast3LB development set with *reduced2* tagset ($\leq$ 40 words)

MODEL	LP	LR	F1	EXACT	POS
Word Only					
Generic	79.61	79.78	79.69	**14.90**	87.29
IG	82.57	82.31	**82.44**	14.24	**91.63**
Gold POS					
Generic	88.08	87.69	**87.89**	**30.59**	100.00
IG	87.56	87.31	87.43	29.61	100.00
Pred. POS					
Generic	85.56	85.38	**85.47**	23.03	96.56
IG	85.32	85.24	85.28	**23.36**	96.56

Table 5: PARSEVAL scores on Cast3LB development set with *reduced3* tagset ($\leq$ 40 words)

3.4 Lemmatization Impact

Being a morphologically rich language, Spanish exhibits a high level of inflection similar to several other Romance languages, for example French and Italian (gender, number, verbal mood). Furthermore, Spanish belongs to the pro-drop family and clitic pronouns are often affixed to the verb and carry functional marks. This makes any small treebank of this language an interesting play field for statistical parsing. In this experiment, we want to use lemmatization as a form of morphological clustering. To cope with the loss of information, we provide the parser with predicted POS. Lemmatization is carried out by the morphological analyzer MORFETTE, (Chrupała et al., 2008) while POS tagging is done by the MElt tagger. Lemmatization performances are on a par with previously reported results on Romance languages (see Table 6)

TAGSET	ALL	SEEN	UNK (13.84%)
baseline	98.39	99.01	94.55
reduced2	98.37	98.88	95.18
reduced3	98.24	98.88	94.23

Table 6: Lemmatization performance on the Cast3LB.

To make the parser less sensitive to lemmatization and tagging errors, we train both tools on a 20 jackkniffed setup[4]. Resulting lemmas and POS tags are then reinjected into the train set. The test corpora is itself processed with tools trained on the unmodified treebank. Results are presented Table 7. They show an overall small gain, compared to the previous experiments but provide a clear improvement on the richest tagset, which is the most difficult to parse given its size (106 tags).

First, we remark that POS tagging accuracy with the baseline tagset when no POS is provided is lower than previously observed. This can be easily explained: it is more difficult to predict POS with morphological information when morphological information is withdrawn from input.

Second, and as witnessed before, reduction of the POS tag sparseness using a simplified tagset and increase of the lexical sparseness by handling OOV words using typographical information have adverse effects. This can be observed in the generic Predicted POS section of Table 7 where the *baseline* tagset is the best option. On the other hand, in IG Predicted POS, using the *reduced3* is better than *baseline* and *reduced2*. Again this tagset is a tradeoff between rich information and data sparseness.

[4]The training set is split in 20 chunks and each one is processed with a tool trained on the 19 other chunks. This enables the parser to be less sensitive to lemmatization and/or pos tagging errors.

TAGSET	LR	LP	F1	EX	POS
Word Only – Generic					
baseline	79.70	80.51	**80.1**	15.23	74.04
reduced2	79.19	79.78	79.48	**15.56**	**89.25**
reduced3	79.92	80.03	79.97	13.16	87.67
Word Only – IG					
baseline	80.67	81.32	**80.99**	**15.89**	75.02
reduced2	80.54	81.3	80.92	15.13	**90.93**
reduced3	80.52	80.94	80.73	15.13	88.53
Pred. POS – Generic					
baseline	85.03	85.57	**85.30**	**23.68**	95.68
reduced2	83.98	84.73	84.35	23.36	**96.78**
reduced3	84.93	85.19	85.06	21.05	96.60
Pred. POS – IG					
baseline	84.60	85.06	84.83	23.68	95.68
reduced2	84.29	84.82	84.55	21.71	**96.78**
reduced3	84.86	85.39	**85.12**	**22.70**	96.60

Table 7: Lemmmatization Experiments

In all cases *reduced2* is below the other tagsets wrt. to Parseval F1 although tagging accuracy is better. We can conclude that it is too poor from an informational point of view.

4 Discussion

There is relatively few works actively pursued on statistical constituency parsing for Spanish. The initial work of Cowan and Collins (2005) consisted in a thorough study of the impact of various morphological features on a lexicalized parsing model (the Collins Model 1) and on the performance gain brought by the reranker of Collins and Koo (2005) used in conjunction with the feature set developed for English. Direct comparison is difficult as they used a different test set (approximately, the concatenation of our development and test sets). They report an F-score of 85.1 on sentences of length less than 40.[5]

However, we are directly comparable with Chrupała (2008)[6] who adapted the Collins Model 2 to Spanish. As he was focusing on wide coverage LFG grammar induction, he enriched the non terminal annotation scheme with functional paths rather than trying to obtain the optimal tagset with respect to pure parsing performance. Nevertheless, using the same split and providing gold POS, our system provides better performance (around 2.3 points better, see Table 8).

It is of course not surprising for a PCFG-LA model to outperform a Collins' model based lexicalized parser. However, it is a fact that, on such small treebank configurations, PCFG-LA are crucially lacking annotated data. It is only by greatly reducing the POS tagset and using either a state-of-the-art tagger or a lemmatizer (or both), that we can boost our system performance.

The sensitivity of PCFG-LA models to lexical data sparseness was also shown on French by Seddah et al. (2009). In fact they showed that performance of state-of-the-art lexicalized parsers (Charniak, Collins models, etc.) were crossing that of Berkeley parsers when the training set contains around 2500–3000 sentences. Here, with around 2,800 sentences of training data, we are probably in a setting where both parser types exhibit similar performances, as we suspect French and Spanish to behave in the same way. It is therefore encouraging to notice that our approach, which relies on accurate POS tagging and lemmatization, provides state-of-the-art performance. Let us add that a similar method, involving only MORFETTE, was applied with success to Italian within a PCFG-LA framework and French with a lexicalized parser, both leading to promising results (Seddah et al., 2011; Seddah et al., 2010).

5 Conclusion

We presented several experiments reporting the impact of lexical sparseness reduction on non lexicalized statistical parsing. We showed that, by using state-of-the-art lemmatization and POS tagging on a reduced tagset, parsing performance can be on a par with lexicalized models that manage to extract more information from a small corpus exhibiting a rich lexical diversity. It remains to be seen whether applying the same kind of simplifications to the rest of the tagset, i.e. on the internal nodes, can further improve parse structure quality. Finally, the methods we presented in this paper are not language specific and can be applied to other languages if similar resources exist.

[5] See `http://pauillac.inria.fr/~seddah/spmrl-spanish.html` for details on comparison with that work.

[6] We need to remove CP and SBAR nodes to be fairly comparable.

TAGSET	MODE	TOKENS	ALL	≤ 70	≤ 40
reduced3	Gen.	pred. POS	83.92	84.27	85.08
	eval. w/o CP/SBAR		*84.02*	*84.37*	*85.24*
baseline	IG	pred. lemma & POS	84.15	84.40	85.26
	eval. w/o CP/SBAR		*84.34*	*84.60*	*85.45*
reduced3	Gen.	gold POS	86.21	86.63	87.84
	eval. w/o CP/SBAR		*86.35*	*86.77*	*88.01*
baseline		gold POS	83.96	84.58	–
	(Chrupała, 2008)				

Table 8: PARSEVAL F-score results on the Cast3LB test set

Acknowledgments

Thanks to Grzegorz Chrupała and Brooke Cowan for answering our questions and making data available to us. This work is partly funded by the French Research Agency (EDyLex, ANR-09-COORD-008).

References

Mohammed Attia, Jennifer Foster, Deirdre Hogan, Joseph Le Roux, Lamia Tounsi, and Josef van Genabith. 2010. Handling unknown words in statistical latent-variable parsing models for arabic, english and french. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Marie Candito and Djamé Seddah. 2010. Parsing word clusters. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

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.

M. Civit and M. A. Martì. 2004. Building cast3lb: A spanish treebank. *Research on Language and Computation*, 2(4):549 – 574.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–69.

B. Cowan and M. Collins. 2005. Morphology and reranking for the statistical parsing of spanish. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 795–802. Association for Computational Linguistics.

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 *Proceedings of PACLIC 2009*, Hong-Kong, China.

Alberto Lavelli and Anna Corazza. 2009. The berkeley parser at the evalita 2009 constituency parsing task. In *EVALITA 2009 Workshop on Evaluation of NLP Tools for Italian*.

Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Probabilistic cfg with latent annotations. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 75–82.

Miguel A. Molinero, Benoît Sagot, and Lionel Nicolas. 2009. A morphological and syntactic wide-coverage lexicon for spanish: The leffe. In *Proceedings of the International Conference RANLP-2009*, pages 264–269, Borovets, Bulgaria, September. Association for Computational Linguistics.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

Slav Petrov and Dan Klein. 2008. Parsing german with latent variable grammars. In *Proceedings of the ACL Workshop on Parsing German*.

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.

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.

Djamé Seddah, Grzegorz Chrupała, Ozlem Cetinoglu, Josef van Genabith, and Marie Candito. 2010.

Lemmatization and statistical lexicalized parsing of morphologically-rich languages. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Djamé Seddah, Joseph Le Roux, and Benoît Sagot. 2011. Towards using data driven lemmatization for statistical constituent parsing of italian. In *Working Notes of EVALITA 2011*, Rome, Italy, December.