Morphology and Reranking for the Statistical Parsing of Spanish

Brooke Cowan
MIT CSAIL
brooke@csail.mit.edu

Michael Collins
MIT CSAIL
mcollins@csail.mit.edu

Abstract

We present two methods for incorporating detailed features in a Spanish parser, building on a baseline model that is a lexicalized PCFG. The first method exploits Spanish morphology, and achieves an F1 constituency score of 83.6%. This is an improvement over 81.2% accuracy for the baseline, which makes little or no use of morphological information. The second model uses a reranking approach to add arbitrary global features of parse trees to the morphological model. The reranking model reaches 85.1% F1 accuracy on the Spanish parsing task. The resulting model for Spanish parsing combines an approach that specifically targets morphological information with an approach that makes use of general structural features.

1 Introduction

Initial methods for statistical parsing were mainly developed through experimentation on English data sets. Subsequent research has focused on applying these methods to other languages. There has been widespread evidence that new languages exhibit linguistic phenomena that pose considerable challenges to techniques originally developed for English; because of this, an important area of current research concerns how to model these phenomena more accurately within statistical approaches. In this paper, we investigate this question within the context of parsing Spanish. We describe two methods for incorporating detailed features in a Spanish parser, building on a baseline model that is a lexicalized PCFG originally developed for English.

Our first model uses morphology to improve the performance of the baseline model. English is a morphologically-impoverished language, while most of the world's languages exhibit far richer morphologies. Spanish is one of these languages. For instance, the forms of Spanish nouns, determiners, and adjectives reflect both number and gender; pronouns reflect gender, number, person, and case. Furthermore, morphological constraints may be manifested at the syntactic level: certain constituents of a noun phrase are constrained to agree in number and gender, and a verb is constrained to agree in number and person with its subject. Hence, morphology gives us important structural cues about how the words in a Spanish sentence relate to one another. The mechanism we employ for incorporating morphology into the PCFG model (the Model 1 parser in (Collins, 1999)) is the modification of its part-of-speech (POS) tagset; in this paper, we explain how this mechanism allows the parser to better capture morphological constraints.

All of the experiments in this paper are carried out using a freely-available Spanish treebank produced by the 3LB project (Navarro et al., 2003). This resource contains around 3,500 hand-annotated trees encoding ample morphological information. We could not use all of this information and adequately train the resulting parameters due to limited training data. Hence, we used development data to test the performance of several models, each incorporating a subset of morphological information. The highest-accuracy model on the development set uses the mode and number of verbs, as well as the number of adjectives, determiners, nouns, and pronouns. On test data, it reaches F1 accuracy of 83.6%/83.9%/79.4% for labeled constituents, unlabeled dependencies, and labeled dependencies, respectively. The baseline model, which makes almost no use of morphology, achieves 81.2%/82.5%/77.0% in these same measures.

We use the morphological model from the aforementioned experiments as a base parser in a second set of experiments. Here we investigate the efficacy of a reranking approach for parsing Spanish by using

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 795–802, Vancouver, October 2005.

arbitrary structural features. Previous work in statistical parsing (Collins and Koo, 2005) has shown that applying reranking techniques to the n-best output of a base parser can improve parsing performance. Applying an exponentiated gradient reranking algorithm (Bartlett et al., 2004) to the n-best output of our morphologically-informed Spanish parsing model gives us similar improvements. Using the reranking model combined with the morphological model raises performance to 85.1%/84.7%/80.2% F1 accuracy for labeled constituents, unlabeled dependencies, and labeled dependencies.

2 Related Work

The statistical parsing of English has surpassed 90% accuracy in the precision and recall of labeled constituents (e.g., (Collins, 1999; Charniak and Johnson, 2005)). A recent proliferation of treebanks in various languages has fueled research in the parsing of other languages. For instance, work has been done in Chinese using the Penn Chinese Treebank (Levy and Manning, 2003; Chiang and Bikel, 2002), in Czech using the Prague Dependency Treebank (Collins et al., 1999), in French using the French Treebank (Arun and Keller, 2005), in German using the Negra Treebank (Dubey, 2005; Dubey and Keller, 2003), and in Spanish using the UAM Spanish Treebank (Moreno et al., 2000). The best-reported F1 constituency scores from this work for each language are 79.9% (Chinese (Chiang and Bikel, 2002)), 81.0% (French (Arun and Keller, 2005), 76.2% (German (Dubey, 2005)), and 73.8% (Spanish (Moreno et al., 2000)). The authors in (Collins et al., 1999) describe an approach that gives 80% accuracy in recovering unlabeled dependencies in Czech.[1]

The project that is arguably most akin to the work presented in this paper is that on Spanish parsing (Moreno et al., 2000). However, a direct comparison of scores is complicated by the fact that we have used a different corpus as well as larger training and test sets (2,800- vs. 1,500-sentence training sets, and 700- vs. 40-sentence test sets).

[1]Note that cross-linguistic comparison of results is complicated: in addition to differences in corpus annotation schemes and sizes, there may be significant differences in linguistic characteristics.

Category	Attributes
Adjective	gender, number, participle
Determiner	gender, number, person, possessor
Noun	gender, number
Verb	gender, number, person, mode, tense
Preposition	gender, number, form
Pronoun	gender, number, person, case, possessor

Table 1: A list of the morphological features from which we created our models. For brevity, we only list attributes with at least two values. See (Civit, 2000) for a comprehensive list of the morphological attributes included in the Spanish treebank.

3 Models

This section details our two approaches for adding features to a baseline parsing model. First, we describe how morphological information can be added to a parsing model by modifying the POS tagset. Second, we describe an approach that reranks the n-best output of the morphologically-rich parser, using arbitrary, general features of the parse trees as additional information.

3.1 Adding Morphological Information

The mechanism we employ for incorporating morphological information is the modification of the POS tagset of a lexicalized PCFG[2] — the Model 1 parser described in (Collins, 1999) (hereafter Model 1). Each POS tagset can be thought of as a particular morphological model or a subset of morphological attributes. Table 1 shows the complete set of morphological features we considered for Spanish. There are 22 morphological features in total in this table; different POS sets can be created by deciding whether or not to include each of these 22 features; hence, there are 2^{22} different morphological models we could have created. For instance, one particular model might capture the modal information of verbs. In this model, there would be six POS tags for verbs (one for each of indicative, subjunctive, imperative, infinitive, gerund, and participle) instead of just one. A model that captured both the number and mode of verbs would have 18 verbal POS tags, assuming three values (singular, plural, and neutral) for the number feature.

The Effect of the Tagset on Model 1 Modifying the POS tagset allows Model 1 to better distinguish

[2]Hand-crafted head rules are used to lexicalize the trees.

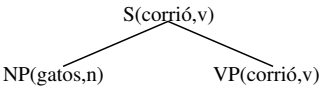

Figure 1: An ungrammatical dependency: the plural noun *gatos* is unlikely to modify the singular verb *corrió*.

events that are unlikely from those that are likely, on the basis of morphological evidence. An example will help to illustrate this point.

Model 1 relies on statistics conditioned on lexical headwords for practically all parameters in the model. This sensitivity to headwords is achieved by propagating lexical heads and POS tags to the nonterminals in the parse tree. Thus, any statistic based on headwords may also be sensitive to the associated POS tag. For instance, consider the subtree in Figure 1. Note that this structure is ungrammatical because the subject, *gatos* (*cats*), is plural, but the verb, *corrió* (*ran*), is singular. In Model 1, the probability of generating the noun phrase (NP) with headword *gatos* and headtag noun (n) is defined as follows:[3]

$$\begin{aligned} P(\text{gatos, n, NP} \mid \text{corrió, v, S, VP}) = \\ P_1(\text{n, NP} \mid \text{corrió, v, S, VP}) \times \\ P_2(\text{gatos} \mid \text{n, NP, corrió, v, S, VP}) \end{aligned}$$

The parser smooths parameter values using backed-off statistics, and in particular smooths statistics based on headwords with coarser statistics based on POS tags alone. This allows the parser to effectively use POS tags as a way of separating different lexical items into subsets or classes depending on their syntactic behavior. In our example, each term is estimated as follows:

$$\begin{aligned} P_1(\text{n, NP} \mid \text{corrió, v, S, VP}) = \\ \lambda_{1,1}\hat{P}_{1,1}(\text{n, NP} \mid \text{corrió, v, S, VP}) + \\ \lambda_{1,2}\hat{P}_{1,2}(\text{n, NP} \mid \text{v, S, VP}) + \\ \lambda_{1,3}\hat{P}_{1,3}(\text{n, NP} \mid \text{S, VP}) \end{aligned}$$

and

$$\begin{aligned} P_2(\text{gatos} \mid \text{n, NP, corrió, v, S, VP}) = \\ \lambda_{2,1}\hat{P}_{2,1}(\text{gatos} \mid \text{n, NP, corrió, v, S, VP}) + \\ \lambda_{2,2}\hat{P}_{2,2}(\text{gatos} \mid \text{n, NP, v, S, VP}) + \\ \lambda_{2,3}\hat{P}_{2,3}(\text{gatos} \mid \text{n}) \end{aligned}$$

[3]Note that the parsing model includes other features such as distance which we omit from the parameter definition for the sake of brevity.

Here the $\hat{P}_{i,j}$ terms are maximum likelihood estimates derived directly from counts in the training data. The $\lambda_{i,j}$ parameters are defined so that $\lambda_{1,1}+\lambda_{1,2}+\lambda_{1,3} = \lambda_{2,1}+\lambda_{2,2}+\lambda_{2,3} = 1$. They control the relative contribution of each level of back-off to the final estimate.

Note that thus far our example has not included any morphological information in the POS tags. Because of this, we will see that there is a danger of the estimates P_1 and P_2 both being high, in spite of the dependency being ungrammatical. P_1 will be high because all three estimates $\hat{P}_{1,1}$, $\hat{P}_{1,2}$ and $\hat{P}_{1,3}$ will most likely be high. Next, consider P_2. Of the three estimates $\hat{P}_{2,1}$, $\hat{P}_{2,2}$, and $\hat{P}_{2,3}$, only $\hat{P}_{2,1}$ retains the information that the noun is plural and the verb is singular. Thus P_2 will be sensitive to the morphological clash between *gatos* and *corrió* only if $\lambda_{2,1}$ is high, reflecting a high level of confidence in the estimate of $\hat{P}_{2,3}$. This will only happen if the context $\langle$corrió, v, S, VP$\rangle$ is seen frequently enough for $\lambda_{2,1}$ to take a high value. This is unlikely, given that this context is quite specific. In summary, the impoverished model can only capture morphological restrictions through lexically-specific estimates based on extremely sparse statistics.

Now consider a model that incorporates morphological information — in particular, number information — in the noun and verb POS tags. *gatos* will have the POS pn, signifying a plural noun; *corrió* will have the POS sv, signifying a singular verb. All estimates in the previous equations will reflect these POS changes. For example, P_1 will now be estimated as follows:

$$\begin{aligned} P_1(\text{pn, NP} \mid \text{corrió, sv, S, VP}) = \\ \lambda_{1,1}\hat{P}_{1,1}(\text{pn, NP} \mid \text{corrió, sv, S, VP}) + \\ \lambda_{1,2}\hat{P}_{1,2}(\text{pn, NP} \mid \text{sv, S, VP}) + \\ \lambda_{1,3}\hat{P}_{1,3}(\text{pn, NP} \mid \text{S, VP}) \end{aligned}$$

Note that the two estimates $\hat{P}_{1,1}$ and $\hat{P}_{1,2}$ include an (unlikely) dependency between the POS tags pn and sv. Both of these estimates will be 0, assuming that a plural noun is never seen as the subject of a singular verb. At the very least, the context $\langle$sv, S, VP$\rangle$ will be frequent enough for $\hat{P}_{1,2}$ to be a reliable estimate. The value for $\lambda_{1,2}$ will therefore be high, leading to a low estimate for P_1, thus correctly assigning low probability to the ungrammatical de-

pendency. In summary, the morphologically-rich model can make use of non-lexical statistics such as $\hat{P}_{1,2}(\mathsf{pn}, \mathsf{NP} \mid \mathsf{sv}, \mathsf{S}, \mathsf{VP})$ which contain dependencies between POS tags and which will most likely be estimated reliably by the model.

3.2 The Reranking Model

In the reranking model, we use an n-best version of the morphologically-rich parser to generate a number of candidate parse trees for each sentence in training and test data. These parse trees are then represented through a combination of the log probability under the initial model, together with a large number of global features. A reranking model uses the information from these features to derive a new ranking of the n-best parses, with the hope of improving upon the baseline model. Previous approaches (e.g., (Collins and Koo, 2005)) have used a linear model to combine the log probability under a base parser with arbitrary features derived from parse trees. There are a variety of methods for training the parameters of the model. In this work, we use the algorithm described in (Bartlett et al., 2004), which applies the large-margin training criterion of support vector machines (Cortes and Vapnik, 1995) to the reranking problem.

The motivation for the reranking model is that a wide variety of features, which can essentially be sensitive to arbitrary context in the parse trees, can be incorporated into the model. In our work, we included all features described in (Collins and Koo, 2005). As far as we are aware, this is the first time that a reranking model has been applied to parsing a language other than English. One goal was to investigate whether the improvements seen on English parsing can be carried across to another language. We have found that features in (Collins and Koo, 2005), initially developed for English parsing, also give appreciable gains in accuracy when applied to Spanish.

4 Data

The Spanish 3LB treebank is a freely-available resource with about 3,500 sentence/tree pairs that we have used to train our models. The average sentence length is 28 tokens. The data is taken from 38 complete articles and short texts. Roughly 27% of the texts are news articles, 27% scientific articles, 14% narrative, 11% commentary, 11% sports articles, 6% essays, and 5% articles from weekly magazines. The trees contain information about both constituency structure and syntactic functions.

Non-Terminal	Significance
aq	*adjective*
cc	*conjunction*
COORD	*coordinated phrase*
ESPEC	*determiner*
GRUP	*base noun phrase*
GV	*verb phrase*
MORF	*impersonal pronoun*
p	*pronoun*
PREP	*base prepositional phrase*
RELATIU	*relative pronoun phrase*
s	*adjectival phrase*
SN	*noun phrase*
SP	*prepositional phrase*
SADV	*adverbial phrase*
S	*sentence*
sps	*preposition*
v	*verb*

Table 2: The non-terminals and preterminals from the Spanish 3LB corpus used in this paper.

4.1 Preprocessing

It is well-known that tree representation influences parsing performance (Johnson, 1998). Prior to training our models, we made some systematic modifications to the corpus trees in an effort to make it easier for Model 1 to represent the linguistic phenomena present in the trees. For the convenience of the reader, Table 2 gives a key to the non-terminal labels in the 3LB treebank that are used in this section and the remainder of the paper.

Relative and Subordinate Clauses Cases of relative and subordinate clauses appearing in the corpus trees have the basic structure of the example in Figure 2a. Figure 2b shows the modifications we impose on such structures. The modified structure has the advantage that the SBAR selects the CP node as its head, making the relative pronoun *que* the headword for the root of the subtree. This change allows, for example, better modeling of verbs that select for particular complementizers. In addition, the new subtree rooted at the S node now looks like a top-level sentence, making sentence types more uniform in structure and easier to model statistically. Additionally, the new structure differentiates phrases em-

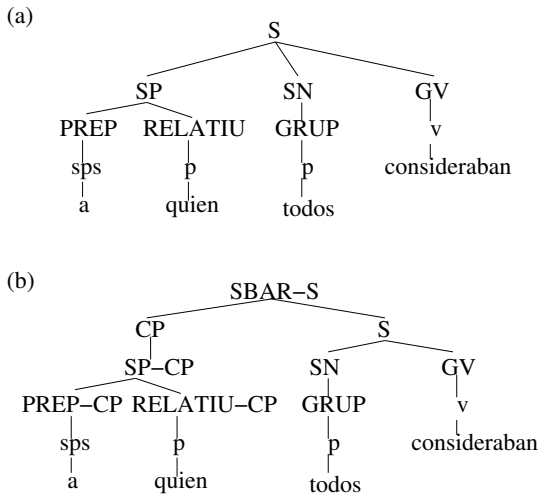

Figure 2: Figure (a) is the original structure from the 3LB treebank for the phrase *a quien todos consideraban* or *whom everyone considered*. We transform structures like (a) into (b) by inserting SBAR and CP nodes, and by marking all non-terminals below the CP with a -CP tag.

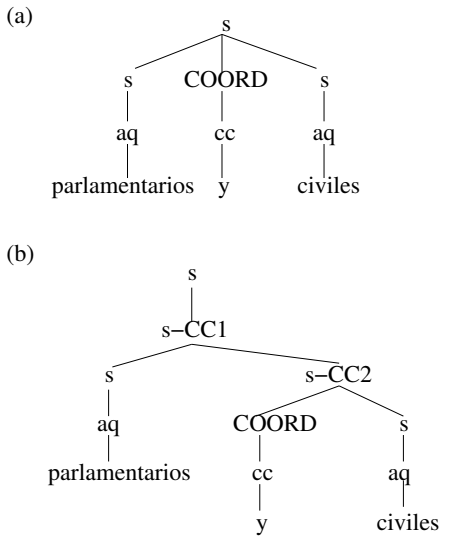

Figure 3: In the 3LB corpus, phrases involving coordination, are represented with a flat structure as in (a). For coordination involving a non-terminal X (X = s in the example), we insert new nodes X-CC1 and X-CC2 to form the structure in (b).

bedded in the complementizers of SBARs from those used in other contexts, allowing relative pronouns like *quien* in Figure 2 to surface as lexical headwords when embedded in larger phrases beneath the CP node.[4]

Coordination In the treebank, coordinated constituents and their coordinating conjunction are placed as sister nodes in a flat structure. We enhance the structure of such subtrees, as in Figure 3. Our structure helps to rule out unlikely phrases such as *cats and dogs and*; the model trained with the original treebank structures will assign non-zero probability to ill-formed structures such as these.

5 Experiments

Our models were trained using a training set consisting of 80% of the data (2,801 sentence/tree pairs, 75,372 words) available to us in the 3LB treebank. We reserved the remaining 20% (692 sentences, 19,343 words) to use as unseen data in a test set. We selected these subsets with two criteria in mind: first, respecting the boundaries of the texts by placing articles in their entirety into either one subset or the other; and second, maintaining, in each subset, the same proportion of genres found in the original set of trees. During development, we used a cross-validation approach on the training set to test different models. We divided the 2,800 training data trees into 14 different development data sets, where each of these data sets consisted of 2,600 training sentences and 200 development sentences. We took the average over the results of the 14 splits to gauge the effectiveness of the model being tested.

To evaluate our models, we considered the recovery of labeled and unlabeled dependencies as well as labeled constituents. Unlabeled dependencies capture how the words in a sentence depend on one another. Formally, they are tuples {*headchild index*, *modifier index*}, where the indices indicate position in the sentence. Labeled dependencies include the labels of the modifier, headchild, and parent non-terminals as well. The root of the tree has a special dependency: {*head index*} in the unlabeled case and {*TOP*, *headchild index*, *root non-terminal*} in the labeled case. The labeled constituents in a tree are all of the non-terminals and, for each, the positions of the words it spans. We use the standard definitions of precision, recall, and F-measure.[5]

[4] This is achieved through our head rules.

[5] When extracting dependencies, we replaced all non-punctuation POS labels with a generic label *TAG* to avoid conflating tagging errors with dependency errors. We also included the structural changes that we imposed during preprocessing. Results for constituent precision and recall were computed after we restored the trees to the original treebank structure.

		Labeled Dep		Unlabeled Dep		Labeled Const			
						<=70 words		<=40 Words	
	Model	**Prec/Rec**	**Gain**	**Prec/Rec**	**Gain**	**Prec**	**Rec**	**Prec**	**Rec**
1	Baseline	76.0	—	82.1	—	81.6	80.4	82.6	81.4
2	n(P,N,V)	78.4	2.4	83.6	1.5	83.1	82.5	84.1	83.4
3	n(A,D,N,P,V)	78.2	2.2	83.5	1.4	83.3	82.4	84.2	83.3
4	n(V)	77.8	1.8	82.9	0.8	82.3	81.6	83.1	82.2
5	m(V)	78.4	2.4	83.1	1.0	82.8	82.0	83.8	82.9
6	t(V)	77.6	1.6	82.7	0.6	82.4	81.4	83.2	82.3
7	p(V)	78.1	2.1	83.3	1.2	82.9	82.0	83.8	82.8
8	g(V)	76.3	0.3	82.2	0.1	81.6	80.6	82.7	81.7
9	n(A,D,N,V,P)+m(V)	**79.0**	**3.0**	**84.0**	**1.9**	**83.9**	**83.2**	**84.7**	**84.1**
10	n(P,N,V)+m(V)	78.9	2.9	83.7/83.8	1.6/1.7	83.6	82.8	84.6	83.7
11	n(A,D,N,V,P)+m(V)+p(V)	78.7	2.7	83.6	1.5	83.6	82.9	84.4	83.8
12	n(A,D,N,V,P)+p(V)	78.4	2.4	83.5/83.6	1.4/1.5	83.3	82.6	84.2	83.5
13	n(A,D,N,V,P)+g(A,D,N,V,P)	78.1	2.1	83.2	1.1	83.1	82.5	83.9	83.4

Table 3: Results after training morphological models during development. When precision and recall differ in labeled or unlabeled dependencies, both scores are shown. Row 1 shows results on a baseline model containing almost no morphological information. The subsequent rows represent a subset of the models with which we experimented: n(P,N,V) uses number for pronouns, nouns, and verbs; n(A,D,N,P,V) uses number for adjectives, determiners, nouns, pronouns, and verbs; n(V) uses number for verbs; m(V) uses mode for verbs; t(V) uses tense for verbs; p(V) uses person for verbs; g(V) uses gender for verbs; the models in rows 9–12 are combinations of these models, and in row 13, n(A,D,N,V,P) combines with g(A,D,N,V,P), which uses gender for adjectives, determiners, nouns, verbs, and pronouns. The results of the best-performing model are in bold.

		Labeled Dep	Unlabeled Dep	Labeled Const			
				<=70 words		<=40 Words	
	Model	**Prec/Rec**	**Prec/Rec**	**Prec**	**Rec**	**Prec**	**Rec**
1	Baseline	77.0	82.5	81.7	80.8	83.1	82.0
2	n(A,D,N,V,P)+m(V)	79.4	83.9	83.9	83.4	85.1	84.4
3	RERANK	80.2	84.7	85.2	85.0	86.3	85.9

Table 4: Results after running the morphological and reranking models on test data. Row 1 is our baseline model. Row 2 is the morphological model that scored highest during development. Row 3 gives the accuracy of the reranking approach, when applied to n-best output from the model in Row 2.

5.1 The Effects of Morphology

In our first experiments, we trained over 50 models, incorporating different morphological information into each in the way described in Section 3.1. Prior to running the parsers, we trained the POS tagger described in (Collins, 2002). The output from the tagger was used to assign a POS label for unknown words. We only attempted to parse sentences under 70 words in length.

Table 3 describes some of the models we tried during development and gives results for each. Our baseline model, which we used to evaluate the effects of using morphology, was Model 1 (Collins, 1999) with a simple POS tagset containing almost no morphological information. The morphological models we show are meant to be representative of both the highest-scoring models and the performance of various morphological features. For instance, we found that, in general, gender had only a slight impact on the performance of the parser. Note that gender is not a morphological attribute of Spanish verbs, and that the inclusion of verbal features, particularly number, mode, and person, generated the strongest-performing models in our experiments.

Table 4 shows the results of running two models on the test set: the baseline model and the best-performing morphological model from the development stage. This model uses the number and mode of verbs, as well as the number of adjectives, determiners, nouns, and pronouns.

The results in Tables 3 and 4 show that adding some amount of morphological information to a parsing model is beneficial. We found, however, that adding more information does not always lead to improved performance (see, for example, rows 11 and 13 in Table 3). Presumably this is because the tagset grows too large.

Table 5 takes a closer look at the performance

of the best-performing morphological model in the recovery of particular labeled dependencies. The breakdown shows the top 15 dependencies in the gold-standard trees across the entire training set. Collectively, these dependencies represent around 72% of the dependencies seen in this data.

We see an extraordinary gain in the recovery of some of these dependencies when we add morphological information. Among these are the two involving postmodifiers to verbs. When examining the output of the morphological model, we found that much of this gain is due to the fact that there are two non-terminal labels used in the treebank that specify modal information of verbs they dominate (infinitivals and gerunds): with insufficient morphological information, the baseline parser was unable to distinguish regular verb phrases from these more specific verb phrases.

Some dependencies are particularly difficult for the parser, such as that in which SBAR modifies a noun ({GRUP TAG SBAR R}). We found that around 20% of cases of this type in the training set involve structures like *el proceso de negociones que* (in English *the process of negotiation that*). This type of structure is inherently difficult to disambiguate. In Spanish, such structures may be more common than in English, since phrases involving nominal modifiers to nouns, like *negotiation process*, are always formed as *noun* + *de* + *noun*.

5.2 Experiments with Reranking

In the reranking experiments, we follow the procedure described in (Collins and Koo, 2005) for creation of a training set with n-best parses for each sentence. This method involves jack-knifing the data: the training set of 2,800 sentences was parsed in 200-sentence chunks by an n-best morphological parser trained on the remaining 2,600 sentences. This ensured that each sentence in the training data had n-best output from a baseline model that was not trained on that sentence. We used the optimal morphological model (n(A,D,N,V,P)+m(V)) to generate the n-best lists, and we used the feature set described in (Collins and Koo, 2005). The test results are given in Table 4.[6]

[6]Note that we also created development sets for development of the reranking approach, and for cross-validation of the single parameter C in approach of (Bartlett et al., 2004).

Dependency	Count	Model	Prec/Rec
Determiner modifier	9680	BL	95.0/95.4
SN GRUP ESPEC L	(15.5%)	M	95.4/95.7
Complement of SP	9052	BL	92.4/92.9
SP PREP SN R	(14.5%)	M	93.2/93.9
SP modifier to noun	4500	BL	83.9/78.1
GRUP TAG SP R	(7.2%)	M	82.9/79.9
Subject	3106	BL	77.7/86.1
S GV SN L	(5.0%)	M	83.1/87.5
Sentential head	2758	BL	75.0/75.0
TOP S	(4.4%)	M	79.7/79.7
S modifier under SBAR	2728	BL	83.3/82.1
SBAR CP S R	(4.4%)	M	86.0/84.7
SP modifier to verb	2685	BL	62.4/78.8
S GV SP R	(4.3%)	M	72.6/82.5
SN modifier to verb	2677	BL	71.6/75.6
S GV SN R	(4.3%)	M	81.0/83.0
Adjective postmodifier	2522	BL	76.3/83.6
GRUP TAG s R	(4.0%)	M	76.4/83.5
Adjective premodifier	980	BL	79.2/80.0
GRUP TAG s L	(1.6%)	M	80.1/79.3
SBAR modifier to noun	928	BL	62.2/60.6
GRUP TAG SBAR R	(1.4%)	M	61.3/60.8
Coordination	895	BL	65.2/72.7
S-CC2 S coord L	(1.4%)	M	66.7/74.2
Coordination	870	BL	52.4/56.1
S-CC1 S-CC2 S L	(1.4%)	M	60.3/63.6
Impersonal pronoun	804	BL	93.3/96.4
S GV MORF L	(1.3%)	M	92.0/95.6
SN modifier to noun	736	BL	47.3/39.5
GRUP TAG SN R	(1.2%)	M	51.7/50.8

Table 5: Labeled dependency accuracy for the top 15 dependencies (representing around 72% of all dependencies) in the gold-standard trees across all training data. The first column shows the type and subtype, where the subtype is specified as the 4-tuple {*parent non-terminal, head non-terminal, modifier non-terminal, direction*}; the second column shows the count for that subtype and the percent of the total that it represents (where the total is 62,372) . The model BL is the baseline, and M is the morphological model n(A,D,N,V,P)+m(V).

5.3 Statistical Significance

We tested the significance of the labeled precision and recall results in Table 4 using the sign test. When applying the sign test, for each sentence in the test data we calculate the sentence-level F1 constituent score for the two parses being compared. This indicates whether one model performs better on that sentence than the other model, or whether the two models perform equally well, information used by the sign test. All differences were found to be statistically significant at the level $p = 0.01$.[7]

[7]When comparing the baseline model to the morphological model on the 692 test sentences, F1 scores improved on 314 sentences, and became worse on 164 sentences. When comparing the baseline model to the reranked model, 358/157 sen-

6 Conclusions and Future Work

We have developed a statistical parsing model for Spanish that performs at 85.1% F1 constituency accuracy. We find that an approach that explicitly represents some of the particular features of Spanish (i.e., its morphology) does indeed help in parsing. Moreover, this approach is compatible with the reranking approach, which uses general features that were first developed for use in an English parser. In fact, our best parsing model combines both the language-specific morphological features and the non-specific reranking features. The morphological features are local, being restricted to dependencies between words in the parse tree; the reranking features are more global, relying on larger portions of parse structures. Thus, we see our final model as combining the strengths of two complementary approaches.

We are curious to know the extent to which a close analysis of the dependency errors made by the baseline parser can be corrected by the development of features tailored to addressing these problems. Some preliminary investigation of this suggests that we see much higher gains when using generic features than these more specific ones, but we leave a thorough investigation of this to future work. Another avenue for future investigation is to try using a more sophisticated baseline model such as Collins' Model 2, which incorporates both subcategorization and complement/adjunct information. Finally, we would like to use the Spanish parser in an application such as machine translation.

Acknowledgements

We would like to thank Xavier Carreras for pointing us to the Spanish 3LB treebank and Montserrat Civit for providing access to the data and answering questions about it. We also gratefully acknowledge the support of the National Science Foundation under grants 0347631 and 0434222.

tences had improved/worse parses. When comparing the morphological model to the reranked model, 199/106 sentences had improved/worse parses.

References

Abhishek Arun and Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: the case of French. ACL 2005, Ann Arbor, MI.

Peter Bartlett, Michael Collins, Ben Taskar, and David McAllester. 2004. Exponentiated gradient algorithms for large-margin structured classification. *Proceedings of NIPS 2004*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine *n*-best parsing and MaxEnt discriminative reranking. ACL 2005, Ann Arbor, MI.

David Chiang and Daniel M. Bikel. 2002. Recovering latent information in treebanks. *Proceedings of COLING-2002*, pages 183–189.

Montserrat Civit Torruella. 2000. Guía para la anotación morfosintáctica del corpus CLiC-TALP. X-Tract Working Paper, WP-00/06.

Michael Collins. 1999. Head-Driven Statistical Models for Natural Language Parsing. University of Pennsylvania.

Michael Collins, Jan Hajic, Lance Ranshaw, and Christoph Tillman. 1999. A statistical parser for Czech. ACL 99.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. EMNLP 2002.

Michael Collins and Terry Koo. 2005. Discriminative Reranking for Natural Language Parsing. *Computational Linguistics*, 31(1):25–69.

C. Cortes and V. Vapnik. 1995. Support Vector Networks. *Machine Learning*, 20:273–297.

Amit Dubey and Frank Keller. 2003. Probabilistic parsing for German using sister-head dependencies. ACL 2003, pp. 96–103.

Amit Dubey. 2005. What to do when lexicalization fails: parsing German with suffix analysis and smoothing. ACL 2005, Ann Arbor, MI.

Mark Johnson. 1998. PCFG Models of Linguistic Tree Representations. *Computational Linguistics*, 24(4):613–632.

Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? ACL 2003, pp. 439–446.

Antonio Moreno, Ralph Grishman, Susana López, Fernando Sánchez, and Satoshi Sekine. 2000. A treebank of Spanish and its application to parsing. *The Proceedings of the Workshop on Using Evaluation within HLT Programs: Results and Trends*, Athens, Greece.

Borja Navarro, Montserrat Civit, Ma. Antònia Martí, Raquel Marcos, and Belén Fernández. 2003. Syntactic, semantic and pragmatic annotation in Cast3LB. *Shallow Processing of Large Corpora* (SProLaC), a Workshop of Corpus Linguistics, 2003, Lancaster, UK.