Resolving prepositional phrase attachment ambiguities in Spanish with a classifier

Nora Aguilar*, Laura Alonso Alemany**, Marina Lloberes*, Irene Castellón*

*Grial Group
Facultat de Filologia, Universitat de Barcelona
Barcelona, Spain
{marina.lloberes,icastellon}@ub.edu

**NLP group
FaMAF – UNC
Córdoba, Argentina
alemany@famaf.unc.edu.ar

Abstract. In this paper we present a classifier that solves a certain kind of ambiguities in syntactic structure for Spanish, namely, ambiguities as to the point of adjunction of a prepositional phrase in the syntactic structure of a sentence (PP attachment).

As a starting point, we used EsTxala dependency grammar for Spanish, integrated within FreeLing, with an accuracy score of 61% on PP adjunction. Our target is to develop a specialized module for for PP attachment, so that the syntactic analyzer combines dependency grammar's manual rules with statistical information infered out of a classifier.

We have evaluated different classifiers and different features to characterize PP-attachment ambiguities. Our best approaches improve the performance of EsTxala by 20 points, but are still far from the performance of unsupervised methods reporting 94% accuracy. We gained insight on the factors governing the disambiguation of PP attachment ambiguities, which will arguably let us build lighter models that can be easily integrated within a general-purpose analyzer as FreeLing.

Keywords: PP attachment, EuroWordNet, empirical methods

1 Introduction and Motivation

Parsing is a fundamental task for NLP, specifically for those applications based on language understanding. The current status of this task for Spanish can still be improved, based in linguistic knowledge or in statistic ones [13, 18, 6, 11, 15, 20]. This paper aims to build a classifier for PP attachment resolution in order to enrich an existing Spanish grammar. This research is part of the enhancement of EsTxala grammar, a rule-based dependency grammar for Spanish developed in the Freeling environment [24].

The problem of PP attachment is that the pattern `VP NP/PP PP`, which humans can parse unambiguously in most of the cases, is always ambiguous for an automatic parser. The following examples show how the same pattern is interpreted as `(VP (NP PP))` in (1) and as `(VP (NP) (PP))` in (2).

(1) John ate the pizza with olives.
(2) The child ate the pizza with a fork.

This syntactic ambiguity is one of the most difficult problems to solve for automatic parsers. That is why specific strategies are usually employed to address it, cascaded on top of the results of general-purpose parsers. Many studies for English [26, 25, 23, 21, 16, 5] show that statistical information about the distribution of prepositions, nouns and verbs does improve the performance of parsing with respect to this problem.

We have focused our research in the resolution of PP attachment for Spanish, to improve the performance of the general-purpose dependency grammar within the FreeLing open-source suite of linguistic analyzers, EsTxala. Current results for EsTxala [19] report about 70% accuracy in head selection (UAS). One of its main problems is precisely PP attachment, which yields about 61% of accurarcy. The final goal of this research is to build a hybrid system, combining symbolic and statistical knowledge for the improvement of parsing.

We base our approach on the assumption that linguistic information is relevant and useful for this task, in contrast with purely textual information, such as the form of words that occur in a `VP PP/NP PP` context. However, it remains to be found which type of linguistic information can help to solve this problem. We expect to assess the contribution of semantic information to this problem. .

The general layout for this article is as follows: In section 2 we present some previous work relevant for this research. In section 3 we describe the EsTxala grammar and discuss its performance with respect to PP attachment. In Section 4 we describe our experimental setting, the manually annotated corpus where we extracted training and test examples from, and the characterization of. We also describe the classifiers we compared for this task. We analyze the results of experiments in Section 5, comparing results for classifiers with baselines and the best performance reported in the literature. Finally, we present some concluding remarks and some lines for future work.

2 Relevant Work

English PP attachment studies can be traced back to Altmann and Steedman (1988) [2], who showed that current discourse context is often useful for disambiguating attachments. Recent work shows that lexical information is generally sufficient [17, 8, 27, 12].

One of the earliest corpus-based approaches to prepositional phrase attachment used lexical preference by computing co-occurence frequencies (lexical asociations) of verbs and nouns with prepositions (Hindle and Roth, 1993) [17]. Training data was obtained by extracting all phrases of the form (`V, N1, P, N2`) from a large parsed corpus.

Ratnaparkhi et al. (1994) [27] used a maximum entropy model considering only lexical information from within the verb phrase (ignoring N2). They experimented with both word features and word class features, their combination yielding 81.6% attachment accuracy.

A non-statistical supervised approach by Brill and Resnik (1994) [8] yielded 81.8% accuracy using a transformation-based approach [7] and incorporating word-class information. Later, Collins and Brooks (1995) [12] achieved 84.5% accuracy by employing a backed-off model to smooth for unseen events. Toutanova et al. (2004) [30] makes use of morphological and syntactic analysis and WordNet synsets, yielding 87.5% accuracy.

An improvement to supervised methods was made via an algorithm that employs a semantically tagged corpus (Stetina and Nagao, 1997) [28]. Each word in a labelled corpus is sense-tagged using an unsupervised word-sense disambiguation algorithm with WordNet [22]. Testing examples are classified using a decision tree induced from the training examples. They report 88.1% attachment accuracy approaching the human accuracy of 88.2% [27].

The unsupervised algorithm of Ratnaparkhi (1998) [26] achieves 81.9% attachment accuracy in English. Using an extraction heuristic, unambiguous prepositional phrase attachments of the form (`V, P, N2`) and (`N1, P, N2`) are extracted from a large corpus. These data model the strength of association of a preposition with noun and verb lemmas. Previously unseen examples of the form `V, N, P, N` are disambiguated by determining whether the preposition is more strongly associated to the noun or to the verb in the example.

Pantel and Lin (2000) [25] describe an unsupervised method that uses a collocation database, a thesaurus, a dependency parser, and a large corpus (125M words), achieving 84.3% precision on Ratnaparkhi's test set.

Studies on PP attachment disambiguation for Spanish are less prolific and more recent than those found for English. The best results are obtained by Ratnaparkhi (1998) [26], who applies his unsupervised method to a journalistic corpus of Spanish, obtaining a 94.5% of accuracy on a test set of 272 examples.

For German, Volk [31] uses the web to obtain n-gram counts, achieving 75% precision. For Spanish, Calvo and Gelbukh (2003) [10] used a variation on Volk's unsupervised method and obtained a 89.5% of coverage, a 91.97% of accuracy and an overall 82.3% when applied to Spanish.

PP attachment disambiguation in Spanish was also carried out within the 2006 CoNLL Shared Task on multilingual dependency parsing. But results do not specify separate results for PP attachment [9].

3 PP attachment resolution in EsTxala

3.1 The EsTxala grammar

EsTxala Dependency Grammar (Lloberes et al., 2010) [19] is an open-source dependency grammar for Spanish implemented in FreeLing environment [14] and developed for FreeLing Dependency Parser module, Txala [3]. EsTxala was designed for providing deeper, robust and wide-coverage parse trees. The current version includes a set of 4,408 rules. In order to deal with these statements, the grammar carries out three basic operations. Input data are constituency partial trees produced by FreeLing Shallow Parser [4], and the grammar builds full syntactic trees and transforms them into dependency trees. After that, each dependency from the tree is labeled with its syntactic function

3.2 The problem of PP attachment in EsTxala grammar

Results of the current EsTxala version [19] yield about 70% accuracy regarding head selection (i.e. unlabeled attachment score). A detailed analysis of results revealed that one of most problematic linguistic phenomena to perform is PP attachment. Currently EsTxala has 61% of PP attachment accuracy. A strategy to improve it would be the inclusion of semantic information in the grammar. However, building a semantic model with verbal and nominal restrictions applied to dependency grammar rules is a complex strategy that doesn't guarantee better results. As a consecuence, we focused on developing a statistical solution in order to improve the results about PP attachment.

4 Experimental Setting

Our experiments are aimed to assess the impact of different kinds of information in the resolution of PP attachment in Spanish.

4.1 The AnCora corpus

We used the Spanish portion of the manually annotated AnCora corpus [29] as provided for CoNLL-2009 shared task. It is constituted by a part of the Lexesp Spanish balanced corpus, the EFE Spanish news agency, and the Spanish version of the newspaper "El Periódico".

In this corpus, we found 4,764 examples of prepositional phrase attachment ambiguities of the form V NP/PP PP. From these examples, 3,171 corresponded to the preposition "de" (or its form "del"), the rest of prepositions were distributed as seen in Table 1, and 46 prepositions (mainly prepositional multiword expressions) occurred only once in the examples.

		prob. of V-attachment		
3171	de	.2	22 contra	3 vía
390	en	.65	19 como	3 frente al
302	a	.63	18 tras	2 salvo
211	para	.62	17 durante	2 respecto a
193	con	.52	15 hacia	2 en vías de
97	por	.55	11 ante	2 en relación a
52	entre	.15	9 a través de	2 en favor de
51	sobre	.27	6 según	2 en cuanto
35	sin	.45	6 frente a	2 debido a
28	desde	.89	4 bajo	2 de acuerdo con
26	hasta	.92	4 apartir de	2 por encima de

Table 1. Number of occurrences of prepositions that occurred more than once in the corpus of examples of the pattern VP NP/PP PP extracted from the AnCora corpus. For the most frequent prepositions, probability of attachment to the verb is provided.

As seen in table 1, the amount of examples for most of the cases is very small. Therefore, one of our aims with this research was to find features that were able to generalize well, even if only few examples were available. We expected that semantic features could provide such power of generalization.

78% of the examples (3,748) were cases were the prepositional phrase was attached to a noun phrase, and the rest (1,015) were attached to a verbal phrase. As displayed in the first column of Table 1 (only for the most frequent prepositions, in the first column), different prepositions presented different probabilities of attachment to the noun or to the verb. For example, in the case of the preposition "de", the most frequent by far, only 1.9% of the examples were attached to the verb. However low the probability, there were an interesting number of 61 examples out of that were attached to the verb, which makes it worthy to try to find a classifier to detect them. But in other frequent cases, the probability of attachment to verb or noun was not so clearly defined. For example, in the case of "en", the second most frequent preposition, in 65% of the examples the prepositional phrase was attached to a verb.

Thus, the simple baseline of attaching a PP to the previous NP can give good results for the preposition "de", but is not useful for the rest of prepositions. Therefore, a more complex approach is needed to deal with the problem of PP attachment.

4.2 Characterization of examples

Examples were characterized by the following features:

- lemma of the preposition dominating the prepositional phrase that has to be attached.
- form and lemma of the preceding noun and preceding verb (4 features).
- number of words from the preposition to the preceding noun and preceding verb (2 features).
- form and lemma of the noun dominated by the preposition (2 features).
- proportion of occurrences of the preposition as a noun or as a verb dependant (2 features). These features were calculated on the development set only.
- proportion of occurrences of the preposition depending from the lemma of the preceding noun and preceding verb (2 features), calculated on the development set.
- concepts from the EuroWordNet Top Concept Ontology [32] that characterize the preceding verb, preceding noun and dominated noun (3 features). Verbs and nouns were not disambiguated, but each lexical item was characterized by every base concept that occurred at least in half of its senses.
- concepts from the EuroWordNet Semantic Field that characterize the preceding verb, preceding noun and dominated noun (3 features). As in the preceding set of features, no disambiguation was carried out.
- whether the preposition was included in the lexical subcategorization frame of the verb (2 features). This information was extracted from the SenSem corpus [1] and is availabe in FreeLing, as files that are used by the dependency grammar.

We carried out different experiments with different subgroups of the above mentioned features, in order to assess the impact of each of these features in the performance of the classifiers. The basic groupings that were evaluated were:

prep preposition lemma (1 feature).

morph form and lemma of preceding noun and verb, number of words to preceding noun and verb (6 features).

morphosynt form and lemma of dominated noun, proportion of dependency from noun, verb, lemma of preceding noun and lemma of preceding verb (6 features). These two last features were left out in most of the experiments because they were very sparse and showed a tendency to overfitting.

syntactic lexical subcategorization either the noun or verb includes the preposition (2 features).

semantic base concepts and semantic file of preceding verb, preceding noun and dominated noun (6 features).

4.3 Experiments

Our goal was to obtain a classifier that used the above mentioned features to decide whether the PP in a previously unseen example of the `VP NP/PP PP` pattern was to be classified as attached either to the noun or to the verb.

In all cases, we trained the classifier with 90% of the examples and tested it with the remaining 10% (477 examples), which had not been seen by the classifier.

Using the weka environment [33], we evaluated different classifiers to gain insight on the resolution of the problem and to assess which could be the best approach. We applied two symbolic classifiers, decision trees (J48) and decision rules (JRip), and two bayesian classifiers, Naive Bayes and Bayes Net.

Additionally, we used three external measures for comparison. First, we used the current performance of EsTxala, described in Section 3.2, is at a 61% of accuracy. We also used two dummy baselines. The most-frequent-class baseline consisted in assigning the most frequent class, adjuncted to the noun, to all examples, which resulted in a 79% accuracy, getting to the level of performance of more complex systems found in the literature. The random baseline consisted in assigning each example one of the two possible classes at random, where the probability of assigning each class was weighted by the probability of occurrence of that class in the corpus of examples. This probability was obtained from the development corpus only. The random baseline performed at 66'3% accuracy, above current EsTxala's performance.

5 Analysis of results

Results obtained by the best-performing subgroups of features are summarized in Table 2, and graphically displayed in Figure 1. Feature sets are described in Section 4.2.

Baselines	
EsTxala	61
Most Frequent Class Baseline	78,7
Random Baseline	66,3

Data-intensive approach	
Ratnaparkhi 1998	94,5

Machine Learning Approaches

morph	morphosynt	synt	sem	prep	lemma-dep	Naive Bayes	Bayes Net	J48	JRip
			+	+		76,94	**85,11**	**86,79**	**87,84**
+				+		**86,16**	**85,11**	84,70	84,28
				+		**85,74**	**85,74**	**85,74**	**85,53**
+	+			+		**85,11**	74,63	19,91	41,92
+	+	+		+		84,70	74,63	19,91	44,02
+	+		+	+		81,97	68,97	19,91	59,53
	+			+		80,71	48,43	19,91	19,91
	+	+		+		80,08	45,28	19,91	19,91
		+				78,62	80,08	80,08	80,08
+			+			66,46	65,83	80,08	78,19
+	+	+				78,19	57,02	19,91	48,64
	+	+	+	+	+	70,86	51,99	19,91	67,92
+	+		+	+	+	70,65	49,06	19,91	19,91
	+		+	+		70,64	50,94	19,91	51,57
+	+		+			69,39	57,23	19,91	48,22

Table 2. Results of different classifiers to decide the point of attachment of a PP in the `VP NP/PP PP` pattern, with different feature sets. Feature sets are ordered by descending accuracy. Results above 85% accuracy are highlighted in boldface.

Figure 1 displays a graphical comparison of the performance of baselines, Ratnaparkhi's unsupervised system and EsTxala with respect to the group of features showing best performance in our approach. It can be seen that the performance of our best approach is closer to the performance of Ratnaparkhi's approach than to any other baseline performance. The performance of Calvo and Gelbukh's approach is comparatively closer to the majority class baseline than to our best approach.

When we compare different subgroups of features within our approach, we see that many of the subgroups perform above the 80% most frequent class baseline. The form of the preposition seems to be the most useful feature for all classifiers, in fact, when only this feature is used, all classifiers perform equally well and not very differently from the best performing approach.

We can see that the higher accuracy score is obtained by the `semantics + preposition` approach, with 87,84% accuracy with JRip and almost 87% with J48. The performance for BayesNet is still as good as 85%, but it drops 10 points for Naive Bayes. This fact seems to indicate that there are complex relationships between semantic features and prepositions, which cannot be captured by the

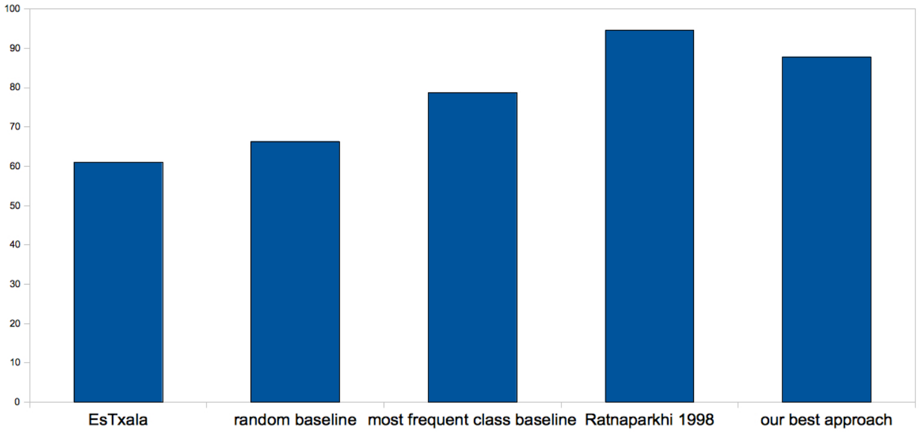

Fig. 1. Performance of different systems that address the problem of PP-attachment: EsTxala, two baselines, Ratnaparkhi's system and our best approach.

simple Naive Bayes model but are indeed captured by the other more complex statistical models.

This indication holds across approaches without semantic information, which produce better results with Naive Bayes, while approaches with semantic information seem to need classifiers that can discover relationships between features. It seems, then, that semantic features can provide a better approximation to the actual causes driving PP-attachment, but this approximation can only be captured by complex relations between features.

The ratio of instances of the preposition depending from the lemma of the preceding noun and preceding verb does not seem to provide useful information. One possible explanation could be that the corpus is not large enough to gather sufficient information for such sparse features as noun and verb lemmas. When evaluated by cross-validation, approaches using these features produce very good results, but they cannot generalize enough to account for previously unseen examples, generally, they overfit to the training examples. Probably, if the development corpus was large enough, these features would provide very valuable information.

This is an interesting result when compared with results obtained for unsupervised approaches, like that of [26]. These approaches rely basically on information about the lexical form of the words co-occurring with a given preposition `VP NP/PP PP` pattern. We believe that the results we obtained with our experiments are in the same line, showing that lexical information (mainly coming from preposition's form) is very useful, but they also show that semantic information can be successfully combined with lexical information, thus complementing unsupervised approaches and providing reliable information whenewer only a small corpus is available.

6 Conclusions and Future Work

We have presented an approach to the problem of PP attachment for Spanish. This approach is to be integrated with the general-purpose grammar in the FreeLing suite of analzyers.

Applying a machine learning approach, we have achieved more than 20% improvement on the performance of the analyzer, and almost 10% improvement over a dummy baseline. We are still far from the 94.5% accuracy reported in the literature for unsupervised approaches, but our model is arguably more compact than one based on the lexical forms of words.

We have carried out an assessment of the impact of different features and different classifiers in the task of PP attachment, and have found that the most useful are the form of the preposition involved in the pattern and also the semantic features of the nouns and verbs involved. Semantic features are very useful because they provide an adequate level of generalization when few examples are available, as is the case of Spanish.

In future work we plan to integrate the results of this research for PP attachment resolution with the EsTxala grammar, in a hybrid approach to parsing. We will also look forward the exploration of the relations of unsupervised approaches with semantic features. We will also utilize some kind of word sense disambiguation, possibly UKB, already included in FreeLing.

Acknowledgements

This work has been partially funded by the Spanish Science and Innovation Ministry, via the KNOW2 project *Language understanding technologies for multilingual domain-oriented information access* (TIN2009-14715-C04), `http://ixa.si.ehu.es/know2`.

Durante el desarrollo de este trabajo, Nora Aguilar contó con una beca del Ministerio de Asuntos de Educación y Ciencia- Agencia Española de Cooperación Internacional para el Desarrollo (MAEC-AECID, 00536854 Programa II.E) así como un complemento de beca para estudios de postgrado en el extranjero de la SecretarÃa de Educación Pública del gobierno mexicano.

References

1. Alonso, L., Capilla, J., Castellón, I., Fernández, A., Vázquez, G.: The sensem project: Syntactico-semantic annotation of sentences in spanish. In: et al., N.N. (ed.) Selected papers from RANLP 2005, pp. 89–98. John Benjamins (2007).
2. Altmann, G., Steedman, M.: Interaction with context during human sentence processing. Cognition 30, 191–238 (1988).
3. Atserias, J., Comelles, E., Mayor, A.: Txala un analizador libre de dependencias para el castellano. Procesamiento del Lenguaje Natural 35, 455–456 (2005).
4. Atserias, J., Carmona, J., Cervell, S., Màrquez, L., Martí, M.A., Padró, L., Placer, R., Rodríguez, H., Taulé, M., Turmo, J.: An environment for morphosyntactic processing of unrestricted spanish text. In: LREC'98. pp. 915–922 (1998).

5. Baldwin, T., Kordoni, V., Villavicencio, A.: Prepositions in applications: A survey and introduction to the special issue. Computational Linguistics 35(2), (2009).
6. Bick, E.: A constraint grammar-based parser for spanish. In: Proceedings of TIL 2006 - 4th Workshop on Information and Human Language Technology (2006).
7. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics 21, (1995).
8. Brill, E., Resnik, P.: A rule-based approach to prepositional phrase attachment disambiguation. In: COLING'94 (1994).
9. Buchholz, S., Marsi, E.: Conll-x shared task on multilingual dependency parsing. In: CoNLL. pp. 149–164 (2006).
10. Calvo, H., Gelbukh, A.F.: Improving prepositional phrase attachment disambiguation using the web as corpus. In: Sanfeliu, A., Ruiz-Shulcloper, J. (eds.) CIARP. Lecture Notes in Computer Science, vol. 2905, pp. 604–610. Springer (2003).
11. Calvo, H., Gelbukh, A.F.: Diluct: An open-source spanish dependency parser based on rules, heuristics, and selectional preferences. In: NLDB. pp. 164–175 (2006).
12. Collins, M., Brooks, J.: Prepositional phrase attachment trhough a backed–off model. In: Proceedings of the 3rd Workshop on Very Large Corpora. (1995).
13. Ferrández, A., Palomar, M., Moreno, L.: Slot unification grammar. In: APPIA-GULP-PRODE. pp. 523–532 (1997).
14. `http://www.lsi.upc.es/~nlp/freeling/`
15. Gamallo, P., González, I.: Una gramática de dependencias basada en patrones de etiquetas. In: Procesamiento del Lenguaje Natural. pp. 315–323 (2009).
16. Girju, R.: The syntax and semantics of prepositions in the task of automatic interpretation of nominal phrases and compounds: A cross-linguistic study. Computational Linguistics 35(2), 185–228 (2009).
17. Hindle, D., Rooth, M.: Structural ambiguity and lexical relations. Computational Linguistics 19, 103–120 (1993).
18. Järvinen, T., Tapanainen, P.: Towards an implementable dependency grammar. CoRR cmp-lg/9809001 (1998).
19. Lloberes, M., Castellón, I., Padró, L.: Spanish freeling dependency grammar. In LREC'10. pp. 693–699 (2010).
20. Marimon, M.: The spanish resource grammar. In LREC'10 (2010).
21. Merlo, P., Ferrer, E.E.: The notion of argument in prepositional phrase attachment. Computational Linguistics 32(3), 341–378 (2006).
22. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., Tengi, R.: Five papers on wordnet. Special Issue of the Intl. J. of Lexicography 3(4), (1991).
23. Olteanu, M., Moldovan, D.: Pp-attachment disambiguation using large context. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. pp. 273–280. ACL (2005).
24. Padró, L., Reese, S., Lloberes, M., Castellón, I.: Freeling 2.1: Five years of open-source language processing tools. In LREC'10. pp. 931–936 (2010).
25. Pantel, P., Lin, D.: An unsupervised approach to prepositional phrase attachment using contextually similar words. In ACL'00 (2000).
26. Ratnaparkhi, A.: Statistical models for unsupervised prepositional phrase attachment. In ACL-36 (1998).
27. Ratnaparkhi, A., Reynar, J., Roukos, S.: A maximum entropy model for prepositional phrase attachment. In: Proceedings of the ARPA Human Language Technology Workshop. pp. 250–255 (1994).
28. Stetina, J., Nagao, M.: Corpus based PP attachment ambiguity resolution with a semantic dictionary. In: Zhou, J., Church, K.W. (eds.) Proceedings of the Fifth Workshop on Very Large Corpora. pp. 66–80. ACL, Beijing, China (1997).

29. Taulé, M., Martí, M., Recasens, M.: Ancora: Multilevel annotated corpora for catalan and spanish. In: LREC'06 (2006).
30. Toutanova, K., Markova, P., Manning, C.D.: The leaf projection path view of parse trees: Exploring string kernels for hpsg parse selection. In: EMNLP 2004 (2004).
31. Volk, M.: Exploiting the www as a corpus to resolve pp attachment ambiguities. In: Proc. of Corpus Linguistics 2001. (2001).
32. Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., Bertagna, F., Alonge, A., Peters, W.: The eurowordnet base concepts and top ontology. Tech. rep., Paris, France, France (1998).
33. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005).