Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial
ISSN: 1137-3601
revista@aepia.org
Asociación Española para la Inteligencia Artificial
España

Castellón, Irene; Alonso Alemany, Laura; Tinkova Tincheva, Nevena
A procedure to automatically enrich verbal lexica with subcategorization frames
Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, vol. 12, núm. 37, 2008, pp. 45-53
Asociación Española para la Inteligencia Artificial
Valencia, España

ARTÍCULO

A procedure to automatically enrich verbal lexica with subcategorization frames

Irene Castellón[1], **Laura Alonso Alemany**[2,3], **Nevena Tinkova Tincheva**[1]

[1]Departament de Lingüística General
UB, España.
{icastellon,nevenatinkova}@ub.edu

[2]FaMAF, UNC, Argentina.
[3]InCO, UdelaR, Uruguay.
alemany@famaf.unc.edu.ar

Abstract

In this paper we introduce a method for automatically assigning subcategorization frames to previously unseen verbs of Spanish, as an aid to syntactical analysis. Since there is not a consensus on the classes of subcategorization frames, we combine supervised and unsupervised learning. We apply clustering techniques to obtain coarse-grained subcategorization classes from an annotated corpus of Spanish, then evaluate these classes and we finally use them to learn a classifier to assign subcategorization frames to the verbs of previously unseen sentences.

Keywords: Natural Language Processing, Lexicon, Subcategorization

1 Introduction

In this paper we introduce a method that combines supervised and unsupervised learning to learn equivalence classes for verbs in Spanish. These equivalence classes group together verbs that present a similar syntactico-semantic behaviour, thus generalizing over the behaviour of particular examples. Such generalization provides a level of granularity in the analysis between coarse-grained part of speech tags (like Noun, Verb, etc.) and the particularity of individual lexical items (like "*comer*", "*beber*", etc.).

This intermediate level of generalization is arguably useful to improve the performance of automatic linguistic analyzers, specially syntactic analyzers, the so-called parsers. Knowing the syntactico-semantic behaviour of the verb allows to determine which syntactic structures are legal for that particular verb, as in the following examples, marked with an asterisk when they are ungrammatical:

a) * Los niños **duermen** sueños tranquilos. /
Los niños **duermen**.
** The children* **sleep** *quiet dreams. /*
The children **sleep**.

b) Los niños **desean** sueños tranquilos. /
* Los niños **desean**.
The children **wish** *quiet dreams. /*
** The children* **wish**.

c) Los niños **sueñan** sueños tranquilos. /
Los niños **sueñan**.
The children dream quiet dreams. /
The children **dream**.

As can be seen in these examples, the verb *dormir* (*sleep*) is only grammatical in intransitive sentences, without a direct object, while the verb *desear* (*wish*) is only grammatical in transitive sentences, with a direct object, ant the verb *soñar* (*dream*) is grammatical in both kinds of syntactical structures. This information associated to a verb is known as *subcategorization frame*.

Also, in most of the cases the syntactico-semantic behaviour of the verb tends to determine an important part of the meaning of important clausal constituents, like the subject, objects, etc: their **semantic role**. A semantic role is a description of the relationship that a constituent plays with respect to the verb in the sentence. The subject of an active sentence is often the agent or experiencer, while objects tend to be patients or benefactive. It is commonly assumed that the semantic role of clausal constituents is determined by the verbal predicate of clauses, as in the following examples:

a) The representatives **accepted** the report from the Academy.

b) The representatives **suffered** the report from the Academy.

c) The representatives **used** the report from the Academy.

The syntactic structure of these clauses is highly similar, in all cases the pattern of phrases is *NounPhrase-Subject – VerbalGroup-Predicate – NounPhrase-DirectObject – PrepositionalPhrase-Ambiguous*. However, the meaning of these phrases is quite different when different verbs the nucleus of the clause: in a, the Subject Noun Phrase is the *benefactive* of the predicate, while in b it is the *patient* and in c it is the *agent*.

Our approach consists in extrapolating the behaviour of known verbs to unknown ones, combining unsupervised and supervised learning techniques. To do that, we first characterize the behaviour of the verbal senses annotated in the SenSem [4] corpus. Then, we apply clustering techniques to generalize the behaviour of these verbal senses, obtaining coarse-grained classes. These classes group together verbs with similar syntactic behaviour, that is, they represent distinct verbal subcategorizations. Each annotated example in the SenSem corpus is assigned to one of these classes. From these tagged examples, we learn a classifier that can assign an unseen example to one of the coarse-grained classes obtained from the corpus.

The rest of the paper is organized as follows. In the following Section we give an overview of related work, then we describe the annotated corpus we learn from and how examples are transformed to represent subcategorization patterns, and the way we have processed it to generalize the learning data. Then, in Section 4 we present our method to create coarse-grained equivalence classes of verbs, and the procedures to evaluate them. In Section 5 we describe some of the solutions that we obtained, and justify their adequacy using the proposed evaluation procedures. The solution found most adequate at this stage of research is further analyzed in Section 6, and we explain some further work we have carried out to refine the classes obtained in the initial clustering solution. Finally, in Section 7 we draw some conclusions and sketch our future work.

2 Related Work

It is commonly assumed that subcategorization frames can significantly improve the performance of automatic syntactic analyzers of natural language. However, the manual construction of lexica with subcategorization information is very costly. That's why there have been several approaches to acquiring such information automatically. A common approach consists in obtaining the structures verbs occur with (the so-called *subcategorization patterns*) in syntactically analyzed corpora. This procedure usually consists in 2 steps: acquiring the subcategorization frame of verbs and then generalizing over the behaviour of particular verbs by finding equivalence classes that group together verbs with similar behaviours. In this work we focus in the second aspect, finding equivalence classes.

Much interesting work has been produced in the last fifteen years in the area of subcategorization acquisition, a good review can be found in [14]. Here we highlight the main differences of our work with respect to some well-established previous work.

A big difference is found in the information provided by the subcategorization patterns of verbs, which is also dependent on the corpus subcate-

corpus is analyzed automatically [13] or not annotated at all [1], in many other cases subcategorizations are acquired from a manually annotated corpus [11, 3]. Different kinds of annotation make it possible to distinguish verbal senses [9] or else it is necessary to work at the level of verb lemma [1, 3], leaving ambiguous verbs as such.

When working with examples from corpus, it is necessary to discriminate which constituent patterns are determined of the verb's subcategorization behaviour, and which are not verb-dependent, that is, which constituents are *arguments* and which are *adjuncts*, respectively. In order to discriminate argumental patterns from patterns with adjuncts, most of the previous work applies filters to the examples from corpus, like a frequency threshold [10], hypothesis testing [1, 5], dismissal of phrases with a given category, etc.

The SenSem corpus provides information about constituents that are arguments in each example, so adjuncts can be discarded to model examples. Since SenSem provides information about verbal senses, our unit is not the verbal lemma, but the verbal sense, because verbal senses can have very differring subcategorization patterns.

With respect to the method for establishing equivalence classes, different approaches have been taken. [2] uses a confidence interval for indicative cues to classify between two classes of verbs, [12] use decision trees and [15] and [7] use a hierarchical clustering algorithm. In this work we use unsupervised clustering using the EM algorithm for clustering. However, as will be seen in the analysis, it seems more adequate to employ a hierarchical clustering algorithm, which we will do in future work.

Last but not least, it is important to note that most of the work in subcategorization acquisition has been done for English. Only a few works can be found for other languages, particularly for Spanish we know of [5, 7].

3 Representing subcategorization from a corpus

3.1 The annotated corpus

Our departure point is SenSem [4], an annotated corpus of Spanish consisting of 25,000 naturally occurring clauses of newspaper text that are tagged with a verbal sense, and where sentence constituents have been annotated with their morphosyntactic category, syntactic function and semantic role. The most frequent 250 verbs of Spanish are represented, and over 1100 senses are distinguished. Each sense in SenSem has been associated to a subcategorization frame obtained as a synthesis of the structures found in the examples of the corpus.

From that corpus, we characterize verbal senses by the arguments they occur with in annotated examples, regardless of the order they occur with. Each verbal sense is characterized as a vector. The vector space for all senses consists of every realization found in the annotated corpus. The value of each vector in each dimension is the number of times that sense has occurred with that particular realization. This kind of representation of verbal behaviour is generally accepted as an adequate representation of the subcategorization frame of verbs. See Figure 1 for an illustration.

Different transformations of the corpus are carried out, thus configuring different spaces, as explained in the following Section.

3.2 Transformations of examples

Examples in the corpus are transformed in order to reduce the attribute space and also the data sparseness problem. In the first place, categories are collapsed as seen in Table 2, thus reducing variation in the realizations to be found, and the order of occurrence of constituents is not taken into consideration.

Then, we consider different subsets of the information available for each example: category of constituents only, category and syntactic function, and finally we also characterize examples with the whole of the available information: category, function and semantic role. Moreover, we also reduce the attribute space by considering only realizations that occur more than 5 or 10 times in the corpus. These different configurations significantly change the size of the attribute space, as can be seen in Table 1, but they also change the detail by which examples are described. Reducing the level of detail is beneficial

	DirObj:NP-Subj:NP	PrepObj:PP-Subj:NP	Subj:NP	DirObj:NP	PrepObj:PP
aclarar_6	26	0	2	2	0
acceder_2	0	70	0	0	5

Figure 1: Illustration of how verbal senses can be characterized in terms of its contexts of occurrence, with a subset of the patterns of realization in the corpus.

sparseness, as is the case when examples are characterized by category, function and semantic role. However, if examples are poorly characterized, reducing the number of attributes may produce a significant information loss.

We have to take into account that some of the information we are using to characterize manually annotated examples will not be available for unseen examples, like for example argumentality, semantic role, or even syntactic function. However, to induce equivalence classes, we resort to some of the information that is available in the manually annotated corpus, so that classes are well founded. Then, the problem of classification will have to deal with the difference between the way subcategorization classes are learned and the information available for previously unseen verbs.

4 Equivalence classes

Then, we apply clustering techniques to obtain classes of verbal senses that are similar according to their realizations in the corpus, that is, verbal senses that have similar subcategorization behaviours. We use some of the clustering algorithms provided by Weka [16]. More specifically, we have tried Simple KMeans [8] and Expectation-Maximization clustering (EM) [6].

EM is specially suited for our purposes because the method can find an optimal number of classes for a given dataset, so that the number of classes is not provided by the researcher as an additional bias. In order to find the optimal clustering, the EM method assumes the cluster points follow certain probability distribution, and so it groups points in clusters that are optimal based on that assumption. Since we use Weka, we are assuming a Gaussian distribution, but we did not check whether the data actually follow that distribution. However, in comparison with Simple KMeans, EM provides results that are linguistically more adequate.

As with all unsupervised techniques, evaluation is an unclear issue. Since we have not implemented this method in a final application, we cannot use the kind of indirect evaluation obtained from the impact in application's performance. However, we have envisaged some methods to help evaluate the adequacy of different clustering solutions.

4.1 Qualitative evaluation

In the first place, a manual, qualitative evaluation of clustering solutions was carried out. We studied the **population of clusters**, and clustering solutions that presented classes with only one verb were dispreferred, because singleton classes do not provide any generalization on the behaviour of the verbs. We also created a list of **pairs of highly similar verb senses**, shown in Table 3, and checked whether they were assigned to the same cluster or to different clusters, the latter being an indicator of bad clustering solutions. Finally, we also inspected the **global content** of clusters, and determined whether the majority of verbs in each cluster actually shared similar subcategorization behaviour (for example, if they were all transitives, ditransitives, etc.).

4.2 Quantitative evaluation

As for objective metrics, we developed two quantitative methods for the intrinsic evaluation of clustering solutions. The metric **overlap** (O) measures the amount of subcategorization patterns that are shared by different clusters, weighted by the relative frequency of each pattern in each cluster, calculated as follows:

$$O_{A,B} = \frac{\sum_{p \in (P_A \cap P_B)} F_A(p) + F_B(p)}{\sum_{p \in (P_A \cup P_B)} F_A(p) + \sum_{p \in (P_B \cup P_A)} F_B(p)} \quad (1)$$

where

A, B clusters
P_A set of patterns p in A
$F_A(p)$ frequency of occurrence of pattern p in A

	all realizations	realizations > 5	realizations > 10
category	240	98	69
category + function	785	213	130
category + function + role	2854	44	317

Table 1: Reduction of the attribute space by using different subsets of the information associated to examples and by discarding unfrequent realizations.

annotation in corpus	compact	annotation in corpus	compact
Pronoun[1]	Noun Phrase	Comparative Phrase (Subject or Direct Object)	Noun Phrase
Pronoun[2]	Prep. Phrase	Comparative Phrase (Indirect or Prep. Object)	Prep. Phrase
Personal Pronoun[1]	Noun Phrase	Prepositional subordinated clause	Prep. Phrase
Personal Pronoun[2]	Prep. Phrase	Prepositional infinitive clause	Prep. Phrase
Relative Pronoun	Noun Phrase	Gerund clause	Adv. Phrase
Relative Clause[1]	Noun Phrase	Adverbial clause	Adv. Phrase
Interrogative Clause	Noun Phrase	clitic pronoun	Noun Phrase
Direct Speech	Noun Phrase	ellided subject	Noun Phrase
Reduced Clause	Noun Phrase	verbal particle	particle
Proper Name	Noun Phrase	negative particle	–

Table 2: Compactation of morphosyntactic categories of annotated constituents (syntactic functions and semantic roles are left unchanged). Categories not shown in this table are not changed.

We assume that low overlap between classes indicates that the classes contain verbal senses with different syntactic behaviours, while a higher overlap indicates that verbs in different classes share an important part of their syntactic behaviour, which is not intended in our case. As can be expected, the index of overlap is conditioned by the number of classes: the more classes, the higher the chances that overlap is low.

In many cases, different verbal senses are distinguished by different subcategorization frames. That is why we provide a measure of how different senses are distributed in clusters, **distribution of senses** (SD), calculated as follows:

$$SD = \frac{1}{\#V} \sum_{v \in V} \frac{\#C(v)}{\#S(v)} \qquad (2)$$

where

V is the set of verb lemmas v
$S(v)$ is the set of senses of v
$C(v)$ is the set of clusters where at least one sense of v is found

This indicator must be considered with some caution, since there are some verbal senses that share the same subcategorization frames. In any case, it is useful to complement the overall perspective of the distribution of senses across clusters.

Finally, we considered **classifier accuracy**, that is, the accuracy that automatic classifier could achieve to classify unseen instances in its most adequate cluster. So, we first obtained a clustering solution, then tagged each example in the training corpus with its corresponding cluster, and finally performed ten-fold cross validation of classifiers, which were trained on 90% of the corpus and then evaluated on the 10% that was left, and this procedure was repeated 10 times with the 10 possible different partitions of the corpus. This measure gives us a good idea of the adequacy of a given clustering solution for automatic analysis, and it doesn't present any additional effort, because no additional evaluation corpus is needed. Classifiers were trained and evaluated with Weka.

5 Overall analysis of clustering solutions

In what follows we describe different clustering solutions obtained, using the evaluation methods described in the previous section. Then, in the following section we describe the solution that we found optimal up to this point of experimentation, that is, the solution using as attributes realizations of constituents characterized by category and syntactic function that occur more than 10 times in the corpus.

In general, solutions with the KMeans method

hallar_3	encontrar_3	*lie_1*
acceder_2	entrar_2	*go_in_2*
crear_1	construir_1	*produce_2*
valer_1	costar_1	*cost_1*
contener_1	constituir_1	*contain_2*

Table 3: Verb senses with highly similar subcategorization patterns, which are expected to be assigned to the same cluster in good clustering solutions.

provided worse results than solutions with EM, most of all regarding the *population of clusters*, producing many singleton classes. This caused significantly worse overlap indices, since solutions had less "real" classes than their EM counterparts. However, even if a smaller number of real classes was obtained, similar verbs were clustered in different classes more often than in EM solutions. That is why we discarded KMeans and focused in solutions obtained with EM.

5.1 Morphosyntactic category

If only morphosyntactic categories are used to characterize arguments in the examples, and only realizations that occur more than 5 or 10 times are taken into account, EM clustering provides solutions where the population is well distributed in medium-sized classes. There are very few differences between the solution with realizations that occur more than 5 times and that with realizations that occur more than 10 times.

As can be seen in Figure 3, there is a light degradation of the performance of all classifiers when less attributes are used, which leads us to believe that it is counterproductive to reduce the number of attributes when little attributes are available.

It is difficult to obtain linguistically sound generalizations of the behaviour of the verbs in these classes, because of the high ambiguity of the realizations described by morphosyntactic category only, so these solutions were not considered for further analysis.

5.2 Adding Syntactic Function

With examples characterized both by the morphosyntactic category and syntactic function of arguments, considering all realizations, EM provides an optimum of 2 classes, which is far too coarse-grained for the purpose of enriching a lexicon. Some of the additional measures give very good results for this solution (similar pairs of verbs clustered together, Figure 2, performance of classifiers, Figure 3) precisely because only two classes are distinguished, so in this case these measures should not be considered positive indicators.

When considering only realizations that occur more than 5 times, a solution in 3 classes is obtained, and a solution with 5 classes is obtained when considering only realizations that occur more than 10 times. As will be seen in Section 6, the solution with realizations occurring more than 10 times provides linguistically sound classes and groups together many pairs of similar verbs with respect to the relatively high number of classes distinguished, so this will be the solution chosen for further analysis and development.

5.3 Adding Semantic Role

With examples characterized by their morphosyntactic category, syntactic function and semantic role of arguments, solutions that take into account realizations occurring more than 5 or 10 times are far better than those using all realizations. It can be seen in Figure 3 that automate classifiers perform better for solutions with realizations that occur more than 5 or 10 times, probably because they suffer less from data sparseness. Also the number and population of clusters is more understandable for these solutions, and pairs of similar verbs are grouped together more often (see Figure 2).

In these solutions we find four classes. The biggest one is populated by verbs with virtually any pattern of constituents but with a clear predominance of intransitive diatheses, explained because of the ellision of some aof the arguments in the actual realizations in corpus, together with purely intransitive verbs. A second class is populated by strongly transitive verbs, with few intrasitive diatheses, and the two smallest classes are

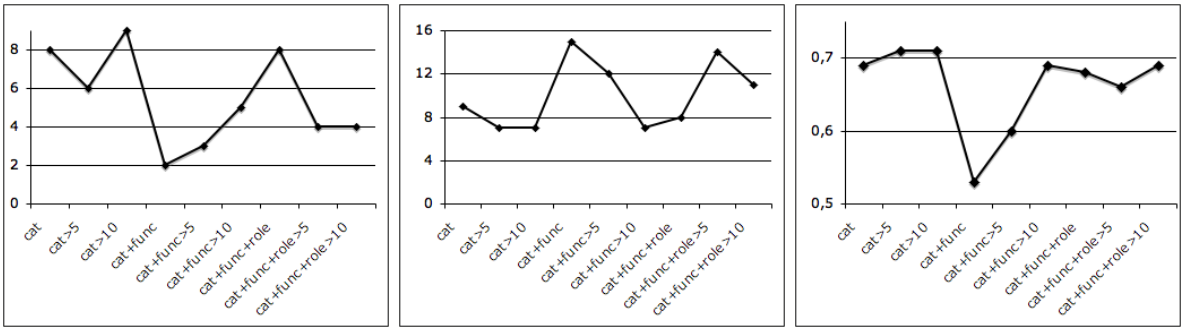

Figure 2: Some objective metrics for comparing clustering solutions: Number of clusters, number of similar verb pairs in the same cluster and distinguishability of senses.

(*origin, goal*), with some intransitive realizations.

These classes were not considered for further analysis because the predominant phenomena (role of intransitive diatheses, transitives, etc.) had already been found in solutions with category and syntactic function only, which is precisely the information that will be available in automatic analysis, so solutions with role were momentarily left aside.

6 Analysis of an interesting clustering solution

As is explained in the previous Section, we chose for further analysis the clustering solution with the EM algorithm provided the most adequate results for our purposes. A detailed description of this solution follows. Then, we explain some further analyses we did to obtain a refinement on the classes of this solution.

6.1 Linguistic description

The chosen clustering solution distinguishes five classes of verb senses, according to their subcategorization patterns:

1. the biggest class, populated with 477 verb senses that alternate beteween **transitive and intransitive** realizations, and some cases of prepositional realizations.

2. a class with 163 senses with predominantly **prepositional and intransitive** realizations. Intransitive realizations can be explained by the omission of the prepositional argument.

3. a class with 103 senses where realizations alternate between **ditransitives, transitives and intransitives**. Realizations with less arguments can mostly be explained by the omission of one or two of the arguments.

4. a class with 68 senses, populated by senses very similar to those in 3.

5. the smallest class, with 63 senses that occur with mostly **prepositional** arguments that alternate with intransitives and some attributes.

It can be seen that these classes contain heterogeneous verbal senses. Therefore, it would be desirable to perform some further clustering within each of these classes to obtain finer-grained distinctions. The optimal way to do that is by applying a hierarchical clustering algorithm, as [15] and [7], but in this first approach we just performed some further EM clustering within each of the classes, in order to inspect their population better.

6.2 Subclustering

When subclustering was performed, EM found no possible distinctions within classes 3 and 4, probably because they are highly compact already. However, eight subclasses were distinguished within class 1: three subclasses group verbal senses with transitive and intransitive realizations, which are most probably transitive verbs that can occur with the ellision of one of its arguments (including the subject). Four other sub-

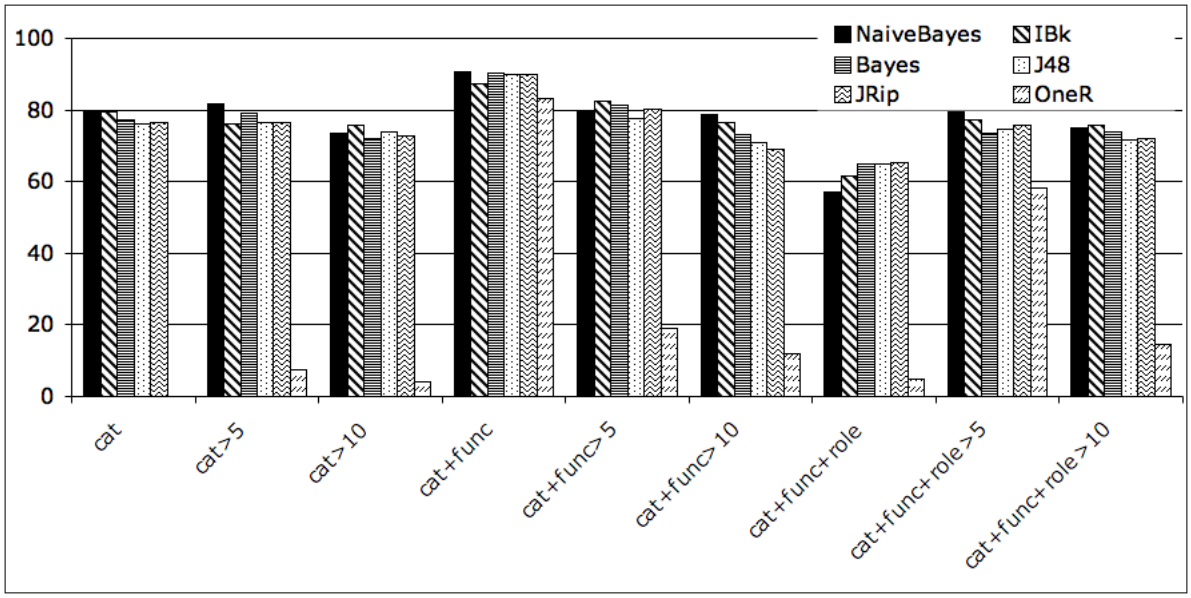

Figure 3: Objective metrics for comparing clustering solutions: classifier accuracy.

classes present prepositional and intransitive realizations, again intransitives can be explained by ellisions. These subclasses are distinguished by the presence in some of them of predicative realizations, transitives, etc.

In class 2 only two subclasses were distinguished, one with ditransitive realizations and the other with realizations with circumstantial arguments. Finally, in class 5, constitued mainly by prepositional realizations, three subclasses are distinguished: one with attributes, another fully prepositional and a third where some transitive realizations can be found.

We can see that at the level of subclasses, it is possible to associate clusters with classical subcategorization frames like *NounPhrase Verb (NounPhrase)* and the like. Therefore, the use of hierarchical techniques seems promising to obtain the granularity of subcategorization information we are looking for, this is left for future work.

We also believe that assessing the contribution of the different features to the constitution of clusters might provide interesting results, both to obtain linguistically sound classes and to increase the accuracy of classifiers.

7 Conclusions

We have presented a procedure to obtain coarse-grained subcategorization classes to assign a subcategorization frame to each verb in a grammar for parsing of Spanish. These classes allow to extrapolate the behaviour of known verbs to unknown verbs, thus providing a procedure to increase the coverage of this kind of information in a grammar at a very low cost.

We have used the information provided in an annotated corpus to characterize the subcategorization behaviour of verbs, then applied clustering techniques to find coarse-grained equivalnece classes of verbs with the same subcategorization behaviour. These classes seem linguistically well motivated and can be automatically recognized with a small error rate. We have developed various methods for evaluating diverse clustering solutions, both qualitatively and quantitatively.

We have found a good clustering solution, that distinguishes verbs with clearly different subcategorization frames. We have performed a further clustering within the classes obtained, and found that a second level is more interesting from the point of view of granularity of the description. Therefore, we will pursue future research by applying hierarchical clustering to the same problem, and we expect to obtain results that provide a more fine-grained, adequate description of verbal senses.

The following phase of this work, once equivalence classes are well established by clustering, is to apply these classes to the same corpus, this time not with manual but with automatic annotation, and evaluate the performance of classifiers in this realistic setting. Then, we will use these classes and the classifier learned from the corpus to assign a subcategorization class to previously unseen verbs, also automatically annotated. We will have to deal with the problem of verb sense disambiguation, and assess how much sense disambiguation contributes to determining the adequate subcategorization frame, and viceversa.

8 Acknowledgements

This research has been partially funded by project KNOW (TIN2006-1549-C03-02) from the Spanish Ministry of Education and Science, a Beatriu de Pinós Postdoctoral Fellowship granted by the Generalitat de Catalunya to Laura Alonso and by a Postgraduate Scholarship FI-IQUC also granted by the Generalitat de Catalunya to Nevena Tinkova, with file number 2004FI-IQUC1/00084.

References

[1] M. R. Brent. From grammar to lexicon. *Computational Linguistics*, 2(19):243–262, 1993.

[2] Michael R. Brent. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 209–214, Morristown, NJ, USA, 1991. Association for Computational Linguistics.

[3] Ted Briscoe and John Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the 35th annual meeting on Association for Computational Linguistics*, 1997.

[4] Irene Castellón, Ana Fernández-Montraveta, Glòria Vázquez, Laura Alonso, and Joanan Capilla. The SENSEM corpus: a corpus annotated at the syntactic and semantic level. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, 2006.

[5] Grzegorz Chrupala. Acquiring verb subcategorization from spanish corpora. Master's thesis, Universitat de Barcelona, 2003.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39, 1977.

[7] Eva Esteve Ferrer. Towards a semantic classification of spanish verbs based on subcategorisation information. In *ACL'04*, 2004.

[8] J. A. Hartigan and M. A. Wong. Algorithm as136: a k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[9] Anna Korhonen. Assigning verbs to semantic classes via wordnet. In *Proceedings of the COLING Workshop on Building and Using Semantic Networks*, Taipei, 2003.

[10] Anna Korhonen, Genevieve Gorrell, and Diana McCarthy. Statistical filtering and subcategorization frame acquisition". In *roceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000.

[11] Anna Korhonen and Judita Preiss. Improving subcategorization acquisition using word sense disambiguation. In *Proceedings of ACL*, pages 48–55, 2003.

[12] Paola Merlo and Suzanne Stevenson. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408, 2001.

[13] A. Sarkar and D. Zeman. Automatic extraction of subcategorization frames for czech. In *COLING'2000*, 2000.

[14] Sabine Schulte im Walde. The Induction of Verb Frames and Verb Classes from Corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook.*, chapter 61. Mouton de Gruyter. To appear.

[15] Sabine Schulte im Walde. Clustering verbs semantically according to their alternation behaviour. In *COLING'00*, pages 747–753, 2000.

[16] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2005.