

# Obtaining coarse-grained classes of subcategorization patterns for Spanish

Laura Alonso i Alemany  
InCo FaMAF  
UdelaR UNC  
Uruguay Argentina  
alemany@famaf.unc.edu.ar

Irene Castellón Masalles Nevena Tinkova Tincheva  
Departament de Lingüística General  
Facultat de Filologia, UB  
Espanya  
icastellon@ub.edu

## Abstract

In this paper we introduce a method for automatically assigning a subcategorization frame to each verb in a grammar for deep parsing of Spanish. Our final objective is to learn a classifier to assign subcategorization frames to previously unseen verbs for which this information is not available in a hand-made lexicon. To do that, we first need to establish classes of equivalence of verbs according to their subcategorization frames. In this paper we describe how we apply clustering techniques to obtain coarse-grained subcategorization classes from an annotated corpus of Spanish and propose a methodology to evaluate them for the application of assigning subcategorization to previously unseen verbs.

## 1 Introduction

In this paper we introduce a method for automatically assigning subcategorization frames to previously unseen verbs of Spanish, as an aid to automated deep parsing. It is commonly believed that this kind of information can significantly improve the performance of automatic parsers.

Our approach consists in extrapolating the behaviour of known verbs to unknown ones. To do that, we first characterize the behaviour of the verbs annotated in the SENSEM [6] corpus. Then, we apply clustering techniques to generalize the behaviour of these verbs, 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.

Our final objective is to apply this classifier to previously unseen verbs. In this paper we focus in the first step, inducing subcategorization classes and evaluating them. [1] presents some experiments on applying these classes to automatically annotated examples.

The rest of the paper is organized as follows. In the following Section 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 2 we present our method to create

coarse-grained equivalence classes of verbs, and the procedures to evaluate them. In Section 3 we describe some of the solutions that we obtained, and justify their adequacy with qualitative linguistic analysis. Finally, in Section 4 we make a quick overview of related work and in Section 5 we draw some conclusions and sketch our future work.

### 1.1 The annotated corpus

Our departure point is SENSEM [6], an annotated corpus of Spanish consisting of 25,000 naturally occurring clauses 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 1161 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, whose dimensions are possible realizations of arguments in a given example. The value of each vector in each dimension is the number of times that sense has occurred with that particular realization. We assume that these realizations are an adequate representation of the subcategorization frame of verbs. See Figure 1 for an illustration. The space of dimensions consists of every realization found in annotated corpus. Different transformations of the corpus are carried out, thus configuring different spaces, as explained in the following Section.

### 1.2 Transformations of examples

We do not work with the examples directly, but we perform a compactation of categories [5], in order to reduce the search space and data sparseness.

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

	Dir.Obj.:NP & Subj.:NP	Prep.Obj.:PP & Subj.:NP	Subj.:NP	Dir.Obj.:NP	Prep.Obj.:PP
<i>aclarar_6</i>	26	0	2	2	0
<i>acceder_2</i>	0	70	0	0	5

**Fig. 1:** Illustration of how verbal senses can be characterized in terms of its contexts of occurrence, with a subset of the patterns of realization that are actually found in the corpus.

	realizations		
	all	> 5	> 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 unfrquent realizations.

<i>hallar_3</i>	<i>encontrar_3</i>	<i>lie_1</i>
<i>acceder_2</i>	<i>entrar_2</i>	<i>go_in_2</i>
<i>crear_1</i>	<i>construir_1</i>	<i>produce_2</i>
<i>valer_1</i>	<i>costar_1</i>	<i>cost_1</i>
<i>contener_1</i>	<i>constituir_1</i>	<i>contain_2</i>

**Table 2:** Verb senses with highly similar subcategorization patterns, which are expected to be assigned to the same cluster in good clustering solutions.

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 for those attribute spaces that suffer from data sparseness, as is the case when examples are characterized by category, function and semantic role. However, for cases where examples are poorly characterized, reducing the number of attributes may produce a significant information loss.

Moreover, 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. To our knowledge, no freely available parser can provide this kind of information reliably for Spanish. However, to induce equivalence classes, we resort to some of the information that is available in the manually annotated corpus, hoping that classes will be better motivated.

## 2 Obtaining 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 [17]. More specifically, we have tried Simple KMeans [10] and Expectation-Maximization clustering (EM) [8].

EM is specially suited for our purposes because the method can find the optimal number of classes for a given dataset, so that the number of classes is not provided by the researcher as an additional bias. For comparison, we also provide some runs with Simple KMeans, but evaluation will show EM is superior.

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, compared with Simple KMeans,

EM results 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.

### 2.1 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. We also found **pairs of highly similar verb senses**, shown in Table 2, and checked whether they were assigned to the same cluster or to different clusters. 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.).

### 2.2 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:

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

where

$A, B$  are clusters

$P_A$  is the set of patterns  $p$  in  $A$

$F_A(p)$  is the frequency of occurrence of pattern  $p$  in  $A$

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, 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:

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

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, since there is no need to develop an additional evaluation corpus. Classifiers were also trained and evaluated with Weka.

### 3 Evaluation 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 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.

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 with realizations occurring more than 10.

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.

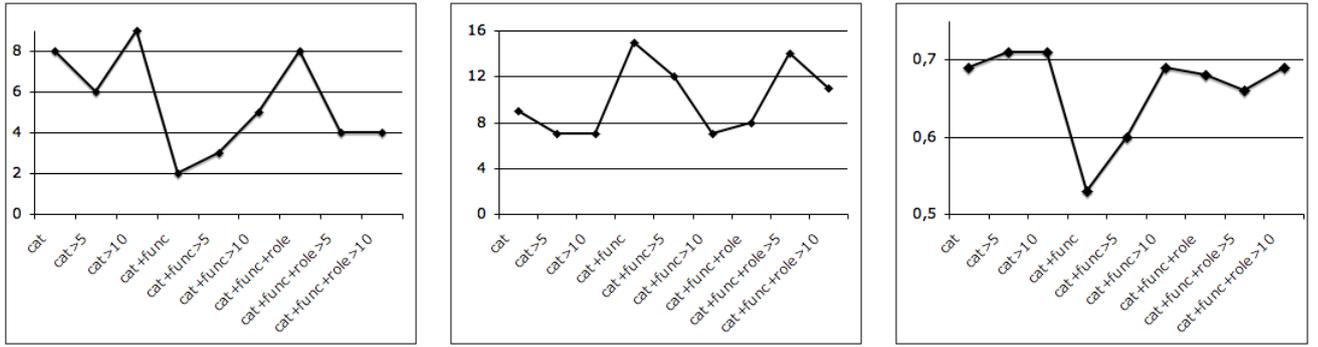
With examples characterized both by the **morphosyntactic category and syntactic function** of arguments, considering all realizations, EM provides an optimum of 2 classes, which is far too coarse-grained for the purpose of enriching a lexicon. Some of the additional measures give very good results for 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 lose their significance.

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 3.1, 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.

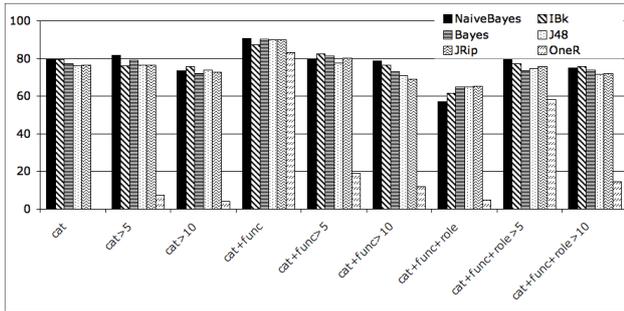
With examples characterized by their **morphosyntactic category, syntactic function and semantic role** of arguments, solutions that take into account realizations occurring more than 5 or 10 times are far better than those using all realizations. It can be seen in Figure 3 that automate classifiers 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 populated by verbs with a very marked semantic roles (*origin, goal*), also with few intransitives.

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.



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



**Fig. 3:** Objective metrics for comparing clustering solutions: classifier accuracy.

### 3.1 Analysis of an interesting clustering solution

We chose for further analysis the clustering solution with the EM algorithm provided the most adequate results for our purposes. Five classes of verb senses are distinguished, according to their subcategorization patterns:

1. the biggest class, populated with 477 verb senses that alternate between **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, we performed some

further clustering within each of these classes to obtain finer-grained distinctions, as described in [5]. We found 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. The optimal way to do that is by applying a hierarchical clustering algorithm, as [16] and [9], but in this first approach we just performed some further EM clustering within each of the classes, in order to inspect their population better. Hierarchical clustering is left for future work.

## 4 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 good review of previous work can be found in [15]. 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 [7, 9]. Here we highlight the main differences of our work with respect to some well-established previous work.

In this work we focus in finding equivalence classes, working upon subcategorization patterns that have already been established in the SENSEM corpus. A big difference is found in the information provided by the subcategorization patterns of verbs, which is also dependent on the corpus subcategorizations are learnt from. In some cases the corpus is analyzed automatically [14] or not annotated at all [3], in many other cases subcategorizations are acquired from a manually annotated corpus [12, 4]. Different kinds of annotation make it possible to distinguish verbal senses [11] or else it is necessary to work at the level of verb lemma [3, 4], leaving ambiguous verbs as such. Since SENSEM provides information about verbal senses, our unit is not the verbal lemma, but the verbal sense.

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. The SENSEM corpus provides information about constituents that are arguments in each example, so adjuncts can be discarded to model examples.

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, [13] use decision trees and [16] and [9] 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.

## 5 Conclusions and Future Work

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

We have used the information provided in an annotated corpus to characterize verbs, then applied clustering techniques to find coarse-grained classes that are 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.

One important line of future work is the use of hierarchical clustering techniques to obtain subcategorization classes at a level of granularity that is more useful for grammatical description. Also as future work, we will use these classes and the classifier learned from the corpus to assign a subcategorization class to previously unseen verbs. 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.

## 6 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] L. Alonso Alemany, I. Castellón, and N. Tinkova Tincheva. Inducción de clases de comportamiento verbal a partir del corpus SENSEM. In *XXIII Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural, XXIII SEPLN*, 2007.
- [2] M. 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] M. R. Brent. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19:243–262, 1993.
- [4] T. Briscoe and J. Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the 35th annual meeting on Association for Computational Linguistics*, 1997.
- [5] I. Castellón, L. Alonso Alemany, and N. Tinkova Tincheva. A procedure to automatically enrich verbal lexica with subcategorization frames. In *Proceedings of the Argentine Symposium on Artificial Intelligence, ASAI'07*, 2007.
- [6] I. Castellón, A. Fernández-Montraveta, G. Vázquez, L. Alonso, and J. 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.
- [7] G. Chrupala. Acquiring verb subcategorization from spanish corpora. Master's thesis, Universitat de Barcelona, 2003.
- [8] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39, 1977.
- [9] E. Esteve Ferrer. Towards a semantic classification of spanish verbs based on subcategorisation information. In *ACL'04*, 2004.
- [10] J. A. Hartigan and M. A. Wong. Algorithm as136: a k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.
- [11] A. Korhonen. Assigning verbs to semantic classes via wordnet. In *Proceedings of the COLING Workshop on Building and Using Semantic Networks*, Taipei, 2003.
- [12] A. Korhonen and J. Preiss. Improving subcategorization acquisition using word sense disambiguation. In *Proceedings of ACL*, pages 48–55, 2003.
- [13] P. Merlo and S. Stevenson. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408, 2001.
- [14] A. Sarkar and D. Zeman. Automatic extraction of subcategorization frames for czech. In *COLING'2000*, 2000.
- [15] S. Schulte im Walde. The Induction of Verb Frames and Verb Classes from Corpora. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook.*, chapter 61. Mouton de Gruyter. To appear.
- [16] S. Schulte im Walde. Clustering verbs semantically according to their alternation behaviour. In *COLING'00*, pages 747–753, 2000.
- [17] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.