# Automatic Acquisition of Syntactic Verb Classes with Basic Resources

L. Mayol, G. Boleda and T. Badia
({laia.mayol,gemma.boleda,toni.badia}@upf.edu)
*GLiCom*
*Dept. of Translation and Philology*
*Pompeu Fabra University*
*Rambla, 30-32*
*08002 Barcelona*
*Tel. +34 93 542 2414*
*Fax: +34 93 542 1617*

**Abstract.**

This paper describes a methodology aimed at grouping Catalan verbs according to their syntactic behavior. Our goal is to acquire a small number of basic classes with a high level of accuracy, using minimal resources. Information on syntactic class, expensive and slow to compile by hand, is useful for any NLP task requiring specific lexical information. We show that it is possible to acquire this kind of information using only a POS-tagged corpus.

We perform two clustering experiments. The first one aims at classifying verbs into transitive, intransitive and verbs alternating with a *se*-construction. Our system achieves an average 0.84 F-score, for a task with a 0.33 baseline. The second experiment aims at further distinguishing among pure intransitives and verbs bearing a prepositional object. The baseline for the task is 0.51 and the upperbound 0.98. The system achieves an average 0.88 F-score.

**Keywords:** lexical acquisition, subcategorisation, verb classes, Catalan, clustering

**Abbreviations:** inf – infinitive; fut – future tense; OBJcli – object clitic; VASE – Verbs alternating with *se*

## 1. Introduction

This paper presents a method to automatically classify Catalan verbs into syntactic classes by means of clustering, an unsupervised machine learning technique. Obtaining lexical information about the linguistic behavior of every word is critical for many NLP tasks, especially in the case of verbs, as they have a great influence on the syntactic pattern and the informational content of the sentence.

However, manually compiling this information is an expensive and slow task, which is never complete and often leads to inconsistent resources (Ide and Véronis, 1998). In the last decade, much research has focused on lexical acquisition, that is, on inferring properties of words from their behavior in corpora and other resources using machine learning techniques.

© 2005 *Kluwer Academic Publishers. Printed in the Netherlands.*

Initial work on automatic acquisition of subcategorisation information was not directed at classifying verbs but at compiling every possible subcategorisation frame for each verb. Brent (1993) used a raw corpus to obtain six different frame types; Manning (1993) described a system which could recognise up to 19 frames; Briscoe and Carroll (1997) followed this same line but dealt with 160 frames. Several works in recent years are closer to the goals or methodology of the experiments presented in this paper. Merlo and Stevenson (2001) applied supervised techniques to acquire three different classes of optionally transitive verbs: unergative, unaccusative and object-drop. They achieved 69.8% accuracy. The technique we use here, clustering, has been used to classify verbs into semantic classes (Schulte im Walde, 2000).

The approach in this and other related work is to use (mainly) syntactic features to induce semantic classes, thus exploiting the syntax-semantics interface. Our task is arguably simpler, because it uses syntactic cues to infer syntactic classes. However, it is by no means trivial, because Catalan syntax is much more flexible than English syntax (Vallduví and Engdahl, 1996) and we use very simple resources, namely, a POS-tagged corpus. If the approach is fruitful, it can be extended to languages with less resources than English or German, such as Catalan itself. The information extracted can be used to create or enhance new resources, such as a parser, and is easy to understand, correct and manipulate by linguists.

Clustering is an unsupervised technique that is used to divide a set of objects into groups. The objects (verbs, in our case) are represented in terms of vectors of features, and the clusters are built based upon the comparison of the feature values. Objects that have similar values are grouped together, and objects with very different values are put in separate clusters. The crucial difference among clustering algorithms is the way they implement the notion of similarity and the clustering procedure (see Kaufman and Rousseeuw (1990) for an overview).

Our motivation for using an unsupervised methodology is twofold: on the one hand, we use it as an empirical test for the classification and set of features. Clustering provides us with insight into the actual structure of the data, and we remodel the classification according to this insight. On the other hand, this methodology facilitates large-scale class induction without need of a large set of manually classified items (contrary to supervised techniques).

The paper is structured as follows: Section 2 introduces the classification; Section 3 explains the materials and methodology of the first experiment, Section 4 its results and Section 5 some extensions; Section 6 discusses the second experiment. Finally, Section 7 presents the conclusions of this paper.

## 2. Classification

Our initial aim was to distinguish between transitive verbs (those subcategorising for an NP object), verbs bearing a prepositional object ("prepositional verbs" from now on), and intransitive verbs (without object of any kind). These classes correspond to the most widely cited distinction in both descriptive and theoretical grammar with respect to verbal syntax. However, the first experimental results made us rethink the classification and the procedure.

We defined a series of features to characterise each class, based on both theoretical and empirical considerations (more details in Section 3.2). When computing two clusters, one corresponded to transitive verbs and the other one to intransitive and prepositional verbs, quite consistent with expectations. However, if more than two clusters were computed, the algorithm made subdivisions of the transitive cluster, and did not separate intransitive from prepositional verbs. The clustering analysis thus signaled that transitive verbs are heterogeneous with respect to their syntactic behaviour, while intransitive and prepositional verbs are not distinct enough.

As for the divisions within transitives, one particular kind of verb tended to be filtered out from more prototypical transitives. These are verbs which require an NP object unless they occur with the particle *se*[1], in which case they require a prepositional object (and admit no NP object), as can be seen in example 1[2]. We call this class VASE (Verbs Alternating with a *SE*-construction).

(1) a. La revolució no **beneficia** tothom
the revolution no benefits everyone
'Revolution doesn't benefit everyone'

b. L'agricultura es **beneficia** del conflicte
the agriculture itself benefits of the conflict
'Agriculture benefits from the conflict'

This class corresponds to an alternation which is very common in Catalan, as well as in other Romance languages (Hernanz and Brucart, 1987; Rosselló, 2002). Due to the quantitative importance of this alternation, and to the fact that these verbs share properties with both transitive and prepositional verbs (they sometimes bear an NP object, sometimes a prepositional one), we added this class to our targeted classification.

As for the problem of distinguishing intransitive from prepositional verbs, we solved it by using a two step procedure. In the first step (Sections 3 to 5), we classify verbs into transitive, intransitive and VASE. Intransitives include

[1] See section 3.2 for a brief overview of the uses of this morpheme in Catalan.
[2] All examples in the paper are taken or adapted from the CTILC corpus (see Section 3.1).

both verbs subcategorising for prepositional objects and pure intransitives[3]. In the second step (Section 6), we distinguish between prepositional verbs and pure intransitives, among all verbs classified as intransitive in the first experiment.

## 3. Experiment 1: Material and method

### 3.1. Data: Corpus and Gold Standard

We used a 16 million word fragment of the CTILC corpus (*Corpus Informatitzat de la Llengua Catalana*; Rafel (1994)). The corpus has been automatically annotated and hand-corrected, providing lemma and morphological information (part of speech and inflectional features).

As mentioned in the previous section, the experiments were carried out on 200 verbs, randomly selected from among those having more than 50 occurrences in the corpus (1288). To be able to evaluate and analyse the results, one of the authors of the paper classified them into the three classes described in the previous section. No agreement scores with other judges were computed, due to the relative straightforwardness of the task. The resulting Gold Standard classification is depicted in Table I.

Table I. Distribution of the Gold Standard across classes.

| Class | # | % |
|---|---|---|
| Transitive | 129 | 64.5 |
| VASE | 39 | 19.5 |
| Intransitive | 32 | 16.0 |

Note that the largest class is by far that of transitive verbs, and that the intransitive class is the smallest one, despite the fact that it includes verbs bearing a prepositional object (12 of the 32 intransitive verbs) and verbs with very infrequent transitive usages (*dormir la migdiada 'take a nap'*, as transitive use of *dormir 'sleep'*).

### 3.2. Features

We designed ten features suitable to characterise the targeted classes, based on literature review and empirical exploration. Each feature was defined in

[3] Apart from experimental considerations, this definition of “intransitive” is in accordance with the Romance linguistics tradition.

terms of superficial linguistic cues which allowed us to automatically extract the data by simple frequency counts. In this section, we describe the features and the expectations we had with respect to their value distributions.

The first three features are directed towards characterising transitive uses of verbs. We therefore expect transitive verbs to have the highest values for these features. VASE verbs should have midrange but nevertheless higher values from those for intransitive verbs, since VASE verbs can occur with an NP object.

**1 ObjCl** Verb cooccuring with an object clitic.

(2) a. No la **veig**
no OBJcli see
'I don't see her'

b. No vull **veure**-la
no want see OBJcli
'I don't want to see her'

**2 NP** Verb followed by a determiner or a noun (that is, an NP), as in sentence (3a). Note that subjects may appear postverbally in Catalan, so that some intransitive verbs may also have relatively high values for this feature. For instance, *aparèixer* 'appear' as in sentence (3b) is intransitive and has this feature. This is a problem especially for unaccusative verbs (such as *aparèixer*), which, following our criteria, should be classified as intransitive but appear more often with postverbal than preverbal subjects. Further features were defined to deal with this problem (see features 7 to 10 below).

(3) a. He **trobat** el meu càstig
have found the my punishment
'I've found my punishment'

b. **Apareixerà** el monstre
Appear-fut the monster
'The monster will appear'

**3 Passive** Verbs appearing in a passive construction, specifically in the following contexts:

- Verb in participle form preceded by the verb *ser* 'to be'.
- Verb in participle form followed by the preposition *per* 'by'.

- Verb preceded by the particle *se* and followed by a noun (corresponding to the so-called ‘passive with *se*’ construction; see example (4b)).

(4) a. Aquest nom fou **proposat** per mademoiselle Scatcherd  
this name was proposed by mademoiselle Scatcherd

b. Es **fregeix** la carn  
SE fries the meat  
‘The meat gets/is fried’

The following two features are expected to capture intransitive uses of verbs, so that transitive and (to a lesser extent) VASE verbs are expected to have lower values for them than intransitive verbs.

**4 Punct** Verb followed by one of the following punctuation marks: full stop, colon, semicolon, exclamation or question mark.

(5) Tecleta **xiscla**.  
Tecleta screams.

**5 Prep** Verb followed by preposition (except for preposition *per* ‘by’; see feature Passive).

(6) Podem **creure** en el miracle  
can believe in the miracle  
‘We can believe in the miracle’

There is only one feature specifically designed to identify VASE verbs:

**6 Se** Verb preceded or followed by the particle *se*, as in example (1b) above. *Se* is a morpheme present in the grammar of most Romance languages, which typically absorbs an argument of the verb. Bartra (2002) recognises five different uses of this particle in Catalan:

- Reflexive particle: it absorbs the object and is roughly equivalent to *itself* in English.
- Possessive dative: it appears with inalienable possession nouns.
- Aspectual pronoun: it establishes the predicate’s telicity.
- Inherent particle: some verbs always appear with a lexically specified *se*.
- Particle in passive constructions: it absorbs the agent.

VASE verbs should have the highest values for this feature and intransitive ones the lowest, since *se* is mostly related to phenomena having to do with transitivity: reflexivity, passivisation, etc. Intransitive (unaccusative) verbs can also co-occur with *se*, as in sentence (7), but we expect these examples to be significantly less frequent.

(7) Quan es **neix**
when himself/herself borns
'When someone is born'

The 6 features listed up to now are represented in terms of raw percentages, that is, their value is obtained by simply dividing the number of sentences where a feature is spotted for a particular verb by the number of occurrences of that verb in the corpus. A different procedure is followed for the last four features, as explained below.

The last four features are aimed specifically at tackling the fact that in Catalan the subject can either be elliptical or appear preceding or following a verb, so that it is not easy to distinguish transitive verbs from intransitive verbs with a postverbal subject. This is a major problem for our task, as explained when discussing feature NP. The same problem would arise for any other language with a similar syntactic pattern, such as Italian or Spanish.

Feature NP may spot a subject or an object. However, it is possible to restrict the contexts in order to ensure with greater certainty that only objects are detected. The following features are elaborations on feature NP which are designed to detect such contexts. Transitive verbs (followed by VASE) verbs should have higher values for these features than intransitive verbs.

**7 2NP** Verb followed by a determiner or a noun and preceded by an adjective, pronoun, determiner[4] or noun. The goal of this feature is to locate a potential object following the verb *and* a potential subject preceding the verb.

(8) Aquest país **consumeix** molt de vinagre
this country consumes much of vinegar
'This country consumes a lot of vinegar'

**8 NonAgrN** Verb followed by a determiner or noun with which it does not agree in number, so that it is much more likely that what follows the verb is the object rather than the subject.

[4] In our POS annotation, a determiner will be tagged as a determiner even if it is functioning as a pronoun.

(9) la nimfa que **contemplà** els balls elegants
the nymph that watched the dances elegant
‘the nymph who watched the elegant dances’

**9 NonAgrP** Verb in first or second person followed by a determiner or a noun, that is, verb followed by a constituent with which it does not agree in person.

(10) Aquí **citem** un fragment de l’ himne
here cite one fragment of the hymn
‘Here we cite a fragment of the hymn’

**10 NonFin** Verb in a nonfinite form followed by a determiner or noun. Subjects are rare with nonfinite verbal forms, though not impossible.

(11) En el moment de **publicar** aquest llibre
in the moment of publish-inf this book
‘When this book was published’

It is easily seen that these last four features are really prone to sparse data problems. Therefore, the values for these features are not computed as raw percentages, but as proportions within the number of times a particular verb occurs with feature NP.

The result of the feature extraction is a representation for each verb as shown in Table II. We see, for example, that 9.3% of the occurrences of the verb *contemplar ‘contemplate’* (transitive) have the feature ObjCl, while *beneficiar ‘benefit’* (VASE) figures in at only 3% for this feature and *xisclar ’scream’* (intransitive) at 0%.

Table II. Feature values for verbs *contemplar*, *beneficiar*, and *xisclar*.

| **Lemma** | **Class** | **ObjCl** | **NP** | **Passive** | **Punct** | **Prep** |
|---|---|---|---|---|---|---|
| *contemplar* | Trans. | 9.3 | 52.2 | 3.4 | 4.3 | 15.0 |
| *beneficiar* | VASE | 3.0 | 20.1 | 2.5 | 6.5 | 32.6 |
| *xisclar* | Intr. | 0 | 11.7 | 0 | 22.0 | 11.0 |

| **Lemma** | **Class** | **Se** | **2NP** | **NonAgrN** | **NonAgrP** | **NonFin** |
|---|---|---|---|---|---|---|
| *contemplar* | Trans. | 5.9 | 15.1 | 17.3 | 13.7 | 25.4 |
| *beneficiar* | VASE | 37.6 | 39.2 | 33.3 | 3.9 | 19.0 |
| *xisclar* | Intr. | 0.8 | 0 | 0 | 0 | 6.6 |

Table III shows the mean values for each feature according to the verb class. Most of the expectations are met: Transitive verbs have the highest values for seven out of the ten features: ObjCl, NP, Passive, 2NP, NonAgrN, NonAgrP and NonFin. Intransitive verbs have the highest values only for Punct and Prep. VASE verbs have values in the midrange for most features (the ones for which transitive verbs have high values, plus Prep), high values for Se and low values for Punct. Some of the differences, such as those for Punct, are not as high as expected, but the patterns are very consistent with our hypotheses.

Table III. Mean values for every feature according to class.

| **Feature** | **Trans.** | **VASE** | **Intr.** |
|---|---:|---:|---:|
| ObjCl | **4.8** | *4.6* | 0.5 |
| NP | **26.4** | *16.3* | 14.1 |
| Passive | **6.5** | *3.1* | 0.6 |
| Punct | *7.1* | 6.8 | **10.9** |
| Prep | 17.3 | *31.3* | **40.2** |
| Se | *11.8* | **33.8** | 2.6 |
| 2NP | **31.9** | *27.6* | 23.0 |
| NonAgrN | **28.4** | *26.5* | 13.2 |
| NonAgrP | **12.4** | *12.2* | 3.1 |
| NonFin | **54.6** | *41.7* | 18.6 |

### 3.3. Clustering approach

For the experiments we used the freely available clustering toolkit CLUTO (Karypis, 2002). We experimented with several algorithms provided in CLUTO (hierarchical and flat, agglomerative and partitional), and the overall structuring of the data was very similar in all approaches. As the results are quite robust, we will only report and analyse the results with the $k$-means algorithm.

$K$-means (see e.g. Kaufman and Rousseeuw (1990)) is a standard flat, partitional clustering algorithm. In its implementation in CLUTO, a first partition into $k$ clusters is randomly computed. In each iteration, the objects in a cluster $i$ are moved to another cluster $j$ if and only if they are closer[5] to the centroid[6] of cluster $j$ than to the centroid of cluster $i$. In each iteration, the centroid is recomputed, and further movements take place if necessary,

[5] We used the cosine as distance measure.

[6] The centroid of a cluster is a vector whose components correspond to the mean values of the components of all objects in the cluster. It is the central point of the multidimensional space of the cluster.

until a user-specified iteration number is reached (20 in our case). $K$-means requires that the number of clusters be defined beforehand. For the sake of clarity, in what follows, we will only report and analyse solutions with three clusters, because our targeted classification consists of three classes.

## 4. Experiment 1: Results and Analysis

### 4.1. Results

As we see in Table IV, cluster 0 contains mainly transitives, cluster 1 intransitives and cluster 2 VASE. Therefore, there is a clear correspondence between classes and clusters, and the cluster analysis has identified the structure we aimed at distinguishing. However, as detailed in the table, there are also some misclassified verbs, which will be further analysed in Section 4.2.

Table IV. Contingency table: Clusters vs. classes in Experiment 1.

| Cluster | Trans. | VASE | Intr. | *Total* |
|---|---|---|---|---|
| 0 | **115** | 7 | 5 | *127* |
| 1 | 9 | 0 | **26** | *35* |
| 2 | 5 | **32** | 1 | *38* |
| *Total* | *129* | *39* | *32* | *200* |

Table V shows the kind of value[7] (high, midrange or low) that each feature has in each cluster. While cluster 0 has the highest values for all features indicating transitivity, cluster 1 has the highest values for the two features that indicate intransitivity. Cluster 2 has midrange values for most features (except for feature Se): It is in between clusters 0 and 1, as VASE are in between transitives and intransitives.

These data fit with the distribution of feature values across classes reported in Table III, showing that the value distribution of the features defined for each class is consistent with the predictions. For example, verbs which have midrange values for features indicating transitivity tend to have a relatively high value for feature Se.

Table VI shows the evaluation measures as compared to the Gold Standard: Precision, recall and F-score. We can use these standard measures because there is a clear correspondence between clusters and classes, that is,

[7] The bands are defined relatively to each feature: For every feature, the cluster with the highest mean value gets ‘High’, the one whose mean falls in the midrange ‘Midrange’, and the one with the lowest mean value ‘Low’, regardless of the absolute value.

Table V. Distribution of feature values.

| **Cluster** | **High** | **Midrange** | **Low** |
|---|---|---|---|
| 0-Trans. | ObjCl<br>NP<br>Passive<br>2NP<br>NonAgrN<br>NonAgrP<br>NonFin | Punct<br>Se | Prep |
| 1-Intr. | Punct<br>Prep | | ObjCl<br>NP<br>Passive<br>2NP<br>NonAgrN<br>NonAgrP<br>NonFin<br>Se |
| 2-VASE | Se | ObjCl<br>NP<br>Passive<br>2NP<br>NonAgrN<br>NonAgrP<br>NonFin<br>Prep | Punct |

we can establish an equivalence between a class *X* (transitive, intransitive or VASE) and the cluster *C* which contains the majority of objects of that class. In that way, precision for *X* is obtained by dividing the number of *X*-verbs in cluster *C* by the total number of elements of the cluster. Recall is computed by dividing the number of *X*-verbs in cluster *C* by the total number of *X*-verbs. F-score is the mean between precision and recall. We use a random baseline. We actually performed a random classification, and computed precision, recall and F-score as explained in this paragraph.

Table VI. Precision, recall and F-score: Clustering results (Cl.) compared to baseline *(Bl.)*.

| | **Prec.** | | **Recall** | | **F-score** | |
|---|---|---|---|---|---|---|
| **Class** | Cl. | *(Bl.)* | Cl. | *(Bl.)* | Cl. | *(Bl.)* |
| Trans. | .91 | *(.62)* | .89 | *(.31)* | .90 | *(.46)* |
| Intr. | .74 | *(.19)* | .81 | *(.41)* | .78 | *(.30)* |
| VASE | .84 | *(.17)* | .82 | *(.28)* | .83 | *(.23)* |
| Average | .83 | *(.32)* | .84 | *(.33)* | .84 | *(.33)* |

The average F-score is 0.84[8] , which can be considered a good overall result for a lexical acquisition task and also when compared to the baseline (0.33). Recall from the Introduction that Merlo and Stevenson (2001) achieved 69.8% accuracy when classifying verbs into unergative, unaccusative and object-drop. Although, in principle, the task presented here is simpler, Catalan syntax is much more flexible than English syntax, which makes feature definition and data extraction more difficult.

Note that the class with the highest score is that of transitives, probably due to the fact that it is the largest class, and most of the defined features are characteristic of transitives, so that the clustering algorithm has richer information to use for them. Conversely, intransitive verbs get the lowest score. The most plausible explanation, apart from it being the smallest class, is that it contains heterogeneous elements: pure intransitives and verbs subcategorising for a prepositional object. Our second experiment will be devoted to that distinction.

## 4.2. Error analysis

In this section, we present the verbs which have been misclassified by our clustering system and discuss possible reasons for this misclassification.

### 4.2.1. *Transitive verbs misclassified into cluster 1-Intr.*

Verb list: *alterar (alter), cessar (dismiss; stop), configurar (set up), consultar (consult), netejar (clean), operar (operate), pensar (think), rectificar (correct), reposar (rest; put again).*

Most of these verbs are either very frequently used without the object (as *netejar* or *operar*) or alternate between an NP and a prepositional object (*cessar de, pensar en*). These verbs are polysemic, and each sense subcategorises for a different frame. For instance, in the ‘stop’ sense *cessar* subcategorises for a prepositional phrase, while in the ‘dismiss’ sense it is a plain transitive. We did not establish a specific class for this alternation and therefore classified this verb as transitive in the Gold Standard (because it subcategorises for an NP in one of the readings). As the ‘stop’ sense is far more frequent, feature values for this verb are closer to intransitive verb values and, accordingly, it is classified in cluster 1.

The ideal treatment for these cases, given our methodology, would have been to treat them as additional classes: just as we treat VASE verbs as a class of verbs that alternates between a transitive use and a SE-construction, we should also consider verbs such as *cessar* as belonging to a class that alternates between a transitive use and a prepositional object construction,

[8] For clarity of exposure, we report and discuss only the average *class* F-score, that is, the mean F-score of the 3 classes. Computing the F-score per token yields 0.87 (note that transitive verbs, the largest group, have the best scores).

and aim at obtaining a cluster for verbs in this alternation class. However, this and other alternations are much less frequent than VASE, so that taking them into account would complicate both the classification and the experimental setup, with relatively little gain. By keeping the classification simple (even at the expence of ignoring some alternations) we can obtain robust results, which can be further used as input in experiments designed to obtain more specific subcategorisation frames.

### 4.2.2. *Transitive verbs misclassified into cluster 2-VASE*

Verb list: *avorrir (bore), coure (cook), errar (be mistaken), espolsar (dust), intensificar (intensify).*

All these verbs appear very frequently with particle *se* in the corpus, most of them due to a causative/noncausative alternation (*El Joan cou la carn,* 'Joan cooks the meat' vs. *La carn es cou* 'The meat gets cooked/cooks'). As the noncausative construction is more frequent, they have values similar to VASE verbs. Again, it would be possible to integrate this alternation in the classification, but it affects a comparatively small number of verbs.

### 4.2.3. *Misclassified intransitive verbs*

Verb list: *concordar (agree)* (classified in cluster 2-VASE); *agradar (like), al.ludir (allude), esmorzar (have breakfast), néixer (be born), regalimar (drip)* (classified in cluster 0-Trans.).

Most mistakes in classifying intransitives are due to idiosyncracies of these verbs. For instance, *esmorzar* and *regalimar* have some transitive uses and *agradar* and *néixer* (unaccusatives) appear in the corpus almost exclusively with a postverbal subject.

### 4.2.4. *Misclassified VASE verbs*

Verb list: *admirar (admire), afegir (add), aprofitar (make the most), compadir (pity), envoltar (surround), servir (serve; be useful), trobar (find).*

All misclassified VASE verbs are in cluster 0-Trans. These errors are due to the fact that the *se* construction of these verbs (i.e. *admirar-se de, aprofitar-se de*) does not often appear in the corpus, so that these verbs have low values for features Se and Prep and, hence, are more similar to transitive verbs than to VASE verbs.

### 4.2.5. *Conclusion*

To sum up, we have seen that the verbs that have been misclassified are in one way or another not **prototypical** within their class. Phenomena such as causativity and unaccusativity have arisen in the error analysis, and also polysemy associated to multiple subcategorisation frames. Verbs affected by these semantic phenomena do not behave as pure transitives, intransitives or VASE would in the syntax. Intuitively, however, they should also not be similar to the prototype of the class in which they have been wrongly placed.

We have analysed the $z$-scores[9] of the verbs, because they indicate how far an item deviates from its distribution's mean, that is, how much further or closer a verb is to the centroid compared to the other objects in the cluster. The intuition is correct for transitive and VASE verbs, but not for intransitive verbs[10]. For two of the clusters, thus, we find that mistakes correspond to distance to the centroid, that is, that verbs which have been wrongly placed in a cluster are not similar to the prototypical verb of the cluster (and therefore, the class). This indicates that cluster analysis can provide a means to approach the notion of prototypicality within a class, although further research is needed to back this hypothesis up. In practical terms, this kind of $z$-score analysis could be used to set a boundary for the manual revision of the $k$ first lemmata on the ranking, those that are furthest from the centroid, when automatically classifying a whole lexicon.

### 4.3. Feature Analysis

Up to now, we have analysed only one solution, namely, that obtained when using all features. The question remains whether all features are equally important or even whether all of them are really necessary for our machine learning system. In order to answer these questions, we performed a set of experiments consisting of subsequently eliminating one feature at a time, so that we could calculate the impact of every individual feature in the results.

If all features were necessary, we would expect that performance would decrease when leaving any of the features out. The results, in Table VII, show that this expectation is only met for five of the features, for which F-scores are worse than using the whole set: Se, Prep, NonFin, NP and 2NP (see left column of the table). The F-score does not decrease when any of the other five features are left out (see right column of the table)[11], and the classifications obtained are almost identical to the solution when using all features.

We take this to mean that all ten features correctly describe our classes (see Table III in Section 3.2, which shows that the value distribution of the features was consistent with our expectations) and can be used in a cluster-

[9] "The $z$-score associated with the $i$th observation of a random variable $x$ is given by $z_i \equiv \frac{x_i - \bar{x}}{\sigma}$, where $\bar{x}$ is the mean and $\sigma$ the standard deviation of all observations $x_1, \ldots, x_n$." Eric W. Weisstein. "z-Score." From *MathWorld*–A Wolfram Web Resource. http://mathworld.wolfram.com/z-Score.htm

[10] 6 out of 12 misclassified verbs in cluster 0-Trans. are in the first 23 (out of 129) positions in the $z$-score rank. All 6 misclassified verbs in cluster 2-VASE are in the first 18 (out of 39) positions. In cluster 1-Intr., misclassified verbs are found in all kinds of positions of the $z$-score ranking.

[11] When leaving feature NonAgrP out, performance even increases a bit, but the difference is only due to 2 misclassified verbs and is not statistically significant.

Table VII. F-score eliminating one feature at a time.

| Removed feature | F-score | Removed feature | F-score |
|---|---|---|---|
| **None** | **0.84** | **None** | **0.84** |
| Se | 0.60 | ObjCl | 0.84 |
| Prep | 0.69 | Passive | 0.84 |
| NonFin | 0.78 | Punct | 0.84 |
| NP | 0.81 | NonAgrN | 0.84 |
| 2NP | 0.83 | NonAgrP | 0.85 |

ing experiment to acquire our target classes. However, five of these features seem to be redundant, as they do not improve the results. We confirmed this hypothesis by clustering the data leaving all 5 features out. The clustering solution was almost identical and the F-score did not change.

Recall that our features, presented in section 3.2, can be divided into four classes: those directed towards (1) identifying transitive uses of verbs, (2) identifying intransitive uses of verbs, (3) identifying VASE verbs, and (4) distinguishing transitive verbs from intransitive verbs with postverbal subject. Our system seems to need only one feature for the first three categories and just two features for the more complex task in (4). In other words, ObjCl and Passive are redundant, since the feature NP already identifies transitive verbs; Punct is redundant, since Prep already identifies intransitive uses; and NonAgrN and NonAgrP are redundant, since NonFin and 2NP already distinguish transitive verbs from intransitive verbs with postverbal subjects. Following this logic, note that if Se, the only feature directed towards characterising VASE verbs, is removed, the results get much worse. Also, intuitively NP and Prep are the two most obvious candidates to be used in identifying transitive and intransitive verbs, respectively. This seems to hold empirically, since the differences of mean values of these features according to class (as seen in Table III) are more robust than for the features that do not improve the results. It would be interesting to find out whether these results can be extended to other languages or whether in a given language the best results are achieved using some of the features redundant for Catalan, given that they do characterise the classes.

## 5. Scaling up the approach

As explained above, the experiment has been performed on 200 randomly chosen lemmata, using a 16 million word corpus with hand-annotation. In order for our system to be reusable, it should also work with more lemmata, automatic tagging, and other kinds of corpora. We discuss this extensions in what follows.

### 5.1. More lemmata and POS-tagger

We performed the experiments on the whole set of 1288 verbs with more than 50 occurrences in the corpus. Our Gold Standard being randomly chosen, it was expected that the results would be quite similar when clustering the 200 verbs of the Gold Standard or the whole set. However, because the number of objects alters the vector space, it was safer to explicitly test for it. The results were equivalent, with a 0.83 mean F-score. This result shows that our Gold Standard was representative enough of all Catalan verbs. The data obtained with this test was used for the second experiment, which will be the topic of the next section.

We replicated the experiments on the same CTILC corpus using a POS-tagger, CatCG (Alsina et al, 2002), instead of the manual annotation. CatCG is a rule-based tagger built by the GLiCom group using Constraint Grammar[12] formalism and tools. Results were similar to those using manual annotation data, F-score being 0.82.

These two tests indicate that the features are robust and not tailored to our Gold Standard (same results when clustering a larger set of objects), and that the methodology does not rely specifically on hand annotation, but on features that can be obtained using a POS-tagger.

### 5.2. Different corpus

Finally, we performed the same experiment on a different corpus: CUCWeb (*Corpus d'Ús del Català a la Web*; Boleda et al. (in preparation)), automatically compiled from the Web [13]. It was obtained by crawling the whole *.es* domain and filtering out Catalan documents with a Naive Bayes language classifier. The 366 million word corpus obtained with this procedure was cleaned up using an additional language filter (based on GLiCom's Catalan lexicon) and a duplicate file detector (20% of the corpus were duplicates). The final version of the corpus contains 125,000 documents and 208 million words, and has been tagged with CatCG (see previous Section).

CUCWeb is 12 times larger than CTILC, but it is not balanced (some topics and genres are overrepresented, some underrepresented), and contains more noise (multilingual documents, spelling mistakes, preprocessing problems due to heterogeneity of formats). The question is whether size will compensate for noise, or, put differently, whether "more data is better data" (Church and Mercer, 1993, 18-19), a slogan that needs principled analysis on

[12] http://www.connexor.com/

[13] Interface to the corpus available at http://catedratelefonica.upf.es/cucweb.

different data (see e.g. Banko and Brill (2001), a case study on confusion set disambiguation that supports the slogan).

We found that on this corpus results decrease to 0.71 F-score. These results are acceptable and still well beyond the baseline, so this test shows that our system can be scaled to other corpora. However, they are 13 points lower than results using CatCG on the manually annotated corpus, a point which deserves attention. Feature analysis shows that the data are indeed much noisier: for most features, the differences between the classes in terms of mean value of the features are less clear. Further exploration of the corpus made it clear that the main source of problems is the type of document. HTML documents typically do not contain one single textual unit, but several, and not only are these units running text, but also menus, name lists, copyright statements, etc. These are not adequately preprocessed, and thus the POS-tagger attempts at analysing units that make no linguistic sense. The resulting tagging contains many more errors than when analysing pure textual input (CTILC corpus). In our case, thus, size does not fully compensate for noise.

## 6. Experiment 2

The goal of the second clustering experiment was to distinguish two classes within the verbs automatically classified in the intransitive cluster by the system:

**Pure intransitives** Verbs which take neither NP nor prepositional object.

**Prepositional verbs** Verbs which require a prepositional object.

In the dataset used for the first experiment, there were only 32 intransitive verbs (see Table I), not enough data to perform a meaningful experiment. Therefore, we used the "intransitive" cluster obtained by clustering the 1288 verbs as explained in Section 5.1, that is, the cluster which contained more intransitive verbs as compared to our Gold Standard. This cluster contained 212 verbs, which were manually classified by one of the authors of the paper into the two targeted classes, pure intransitives and prepositional verbs.

In some cases, it was not obvious whether a verb often followed by a prepositional phrase should be classified as a pure intransitive or as a prepositional verb, because the prepositional phrase could be treated either as an adjunct or as a prepositional object, as in the sentences in example (12).

(12) a. arriba al nostre país
arrives to the our country
'(he/she) arrives to our country'

b. rodolà per les escales
rolled by the stairs
'It rolled down the stairs'

To decide whether these prepositional phrases are arguments or adjuncts we used a test usually used to establish the telicity of a proposition. Telic propositions admit delimiting modifiers (*in ten minutes*) while atelic ones admit durative modifiers (*for ten minutes*). If a verb followed by a prepositional phrase denotes a telic situation, we consider the meaning of the phrase to be inherent in the verb and, therefore, we treat it as a prepositional verb. Otherwise, we will consider the verb a pure intransitive. Following this criteria, *arribar* is classified as a prepositional verb and *rodolar* as a pure intransitive, as we can see in (13).

(13) a. arriba al nostre país en/*durant cinc minuts
arrives to the our country in/*for five minutes

b. rodolà per les escales *en/durant cinc minuts
rolled by the stairs *in/for five minutes

The "intransitive" cluster included also some transitives, given that the clustering solution contained some misclassified transitive verbs. This will affect the upperbound for the task, as discussed below. The results of the manual classification are summarised in Table VIII.

Table VIII. Distribution of the Gold Standard in classes.

| **Class** | **#** | **%** |
|---|---:|---|
| Pure Intr. | 96 | 45.3 |
| Prep. Verb | 107 | 50.5 |
| Transitive | 9 | 4.2 |

Four features were defined to distinguish pure intransitives from prepositional verbs. Two of these features had already been used for Experiment 1 (see Section 3.2): Punct and NP. Pure intransitives are expected to show higher values than prepositional verbs for both Punct and NP.

Two new features were established for this task:

**MostFreqPrep1** This feature corresponds to the percentage of the most frequent preposition following a verb, computed among all prepositions following it. Pure intransitives are expected to show more preposition dispersion and, hence, to have a low value for this feature. In contrast, prepositional verbs are expected to be followed mainly by a single preposition and, therefore, to have a high value.

**MostFreqPrep2** It is defined like MostFreqPrep1 but computing the second most frequent preposition. This feature is aimed at detecting verbs that subcategorise for two different prepositions.

Table IX shows mean values for each feature and for each class, which again meet most of the expectations. Prepositional verbs have higher values for MostFreqPrep1 and MostFreqPrep2; Pure intransitives for Punct and transitives (misclassified verbs) for NP.

Table IX. Mean values for prepositional verbs, pure intransitives and transitives

| **Feature** | **Prep. V.** | **Pure intr.** | **Trans.** |
|---|---|---|---|
| MostFreqPrep1 | **35.2** | 10.7 | *13.1* |
| MostFreqPrep2 | **9.8** | 5.7 | *7.1* |
| Punct | 5.6 | **13.3** | *6.2* |
| NP | 12.9 | *17.2* | **21.3** |

We used the same clustering methodology as in experiment 1 (see Section 3.3). The results of the two cluster solution are detailed in Table X. Cluster 0 contains mainly prepositional verbs and a few transitives and pure intransitives, while cluster 1 contains mainly pure intransitives and a few transitives and prepositional verbs. Again, the clustering results parallel the targeted classification.

Table X. Contingency table: Clusters vs. classes.

| Cluster | Pure Intr. | Prep. V. | Trans. | *Total* |
|---|---|---|---|---|
| 0 | 13 | **99** | 3 | *115* |
| 1 | **83** | 8 | 6 | *97* |
| *Total* | *96* | *107* | *9* | *212* |

Table XI shows the results (precision, recall and F-score) for these experiments. It is equivalent to Table VI above, using a random baseline again. The only difference is that we specify the upperbound for the task, given that 9 objects will be misclassified no matter which cluster they fall into, because they are transitive. The mean F-score, 0.88, is only 10 points away from the upperbound.

Feature values show a symmetrical distribution over our two clusters, as shown in table XII. Features MostFreqPrep1 and MostFreqPrep2 have high mean values for elements in cluster 0 and low mean values for elements in

Table XI. Precision, recall and F-score results. Clustering results (Cl.) compared to baseline *(Bl.)*

| | **Prec.** | | **Recall** | | **F-score** | |
|---|---|---|---|---|---|---|
| **Class** | Cl. | *(Bl.)* | Cl. | *(Bl.)* | Cl. | *(Bl.)* |
| Prep. Verbs | .86 | *(.48)* | .86 | *(.54)* | .86 | *(.51)* |
| Pure Intr. | .86 | *(.52)* | .93 | *(.5)* | .89 | *(.51)* |
| Average | .86 | *(.5)* | .89 | *(.52)* | .88 | *(.51)* |
| Upperbound | .96 | | 1 | | .98 | |

cluster 1. NP and Punct show the opposite pattern. These facts match the feature mean values for each class as shown in Table IX.

Table XII. Distribution of feature values.

| **Cluster** | **High** | **Low** |
|---|---|---|
| 0-Prep. Verbs | MostFreqPrep1<br>MostFreqPrep2 | Punct<br>NP |
| 1-Pure Intr. | Punct<br>NP | MostFreqPrep1<br>MostFreqPrep2 |

As in the previous experiment, misclassified verbs are verbs whose behavior is closer to the behavior of the verbs of the other class:

**Misclassified pure intransitives** Verbs that appear very frequently with a particular kind of locative adjunct and, therefore, show high values for MostFreqPrep1. For instance, *conduir per 'drive on', rodolar per 'roll down/over', xocar contra 'crash into'*.

**Misclassified prepositional verbs** These are verbs that have some transitive uses (*pujar 'go up, raise', baixar 'go down, lower'*) or that very often appear without the prepositional object (*jugar 'play', protestar 'protest'*).

## 7. Conclusions and future work

We have presented a cluster analysis aimed at classifying Catalan verbs according to basic syntactic patterns using very simple resources (a POS-tagged corpus).

We have presented two experiments. In the first one, verbs are classified into transitive, intransitive and verbs alternating with a *se*-construction. We have defined ten features with their associated shallow cues, which are linguistically motivated and which our experiments have empirically validated. We achieve a mean F-score of 0.84 for an experiment whose baseline is 0.33, a good result for a lexical acquisition task. Feature analysis has revealed that the same results can be achieved using only 5 features. We have also shown that the experiment can be scaled up to deal with more lemmata, automatic tagging and different kind of corpora.

The second experiment aims at distinguishing pure intransitive verbs from those subcategorising for a prepositional object. We re-use two features and define two more, achieving 0.88 mean F-score for a task with a 0.51 baseline and 0.98 upperbound.

Our results indicate that it is possible to successfully infer this kind of lexical information for languages with much less resources than English. We believe that our system can be straightforwardly extended to other Romance languages, because it exploits characteristics of Catalan which are also typical of other languages in this family, such as being pro-drop, having rich verbal inflection and allowing postposition of the subject. As shown in this paper, a relatively small, POS-tagged corpus suffices for the classification. This kind of resource already exists for most Romance languages.

Most of the mistakes made by the clustering are also linguistically motivated. Misclassified verbs are those that have some special property (belong to a subclass, present a particular alternation, etc.). This is the most pressing challenge for future work. As our error analysis has shown, other alternations and subclasses should be taken into account. Another major enhancement would be to acquire subcategorisation frames with more than one object (for instance, NP plus dative object). As the complexity of the information to be gathered grows, it will become impossible to avoid the polysemy issue, as we have done in this paper. Most medium to high frequency verbs have more than one sense associated to more than one subcategorisation frame, and this should be properly assessed and evaluated. A promising methodology that addresses this problem has been developed by Korhonen et al (2003).

## Acknowledgements

Many thanks to Toni Martí and Enric Vallduví and all the colleagues from the GLiCom for their useful comments. Special thanks are due to the Institut d'Estudis Catalans for lending us the research corpus, and to Nadjet Bouayad and Sebastian Padó for a critical revision of a previous version of this paper. Also thanks to Tom Rozario for language revision. This work is supported by the Departament d'Universitats, Recerca i Societat de la Informació (grants 2003FI-00867 and 2001FI-00582), and by the Fundación Caja Madrid.

## References

À. Alsina, T. Badia, G. Boleda, S. Bott, À. Gil, M. Quixal, and O. Valentín. 2002. CATCG: a general purpose parsing tool applied. In *Proceedings of Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

M. Banko and E. Brill. 2001. Scaling to Very Very Large Corpora for Natural Language Disambiguation In *Proceedings of ACL 2001*, pages 26–33, Toulouse, France.

A. Bartra. 2002. La passiva i les construccions que s'hi relacionen. In Joan Solà (ed.), *Gramàtica del Català Contemporani*, pages 2111–2179. Empúries, Barcelona.

G. Boleda., S. Bott, R. Meza, C. Castillo, T. Badia, V. López. In preparation. CUCWeb: A Catalan Corpus built from the Web.

M. Brent. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262.

T. Briscoe and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, Washington, USA.

K.W. Church and R.L. Mercer. 1993. Introduction to the special issue on Computational Linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

M.L. Hernanz and J.M. Brucart. 1987. *La Sintaxis*. Crítica, Barcelona.

N. Ide and J. Véronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):1–40.

G. Karypis, 2002. *CLUTO: A Clustering Toolkit*. CLUTO 2.0 user manual.

L. Kaufman and P. J. Rousseeuw. 1990. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, New York City, NY.

A. Korhonen, Y. Krymolowswski, and Z. Marx. 2003. Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings the 41st Annual Meeting of the Association for Computational Linguistics*, pages 64–71, Sapporo, Japan.

C. Manning. 1993. Automatic acquisition of a large subcategorisation dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242, Columbus, USA.

P. Merlo and S. Stevenson. 2001. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3):373–408.

J. Rafel. 1994. Un corpus general de referència de la llengua catalana. *Caplletra*, 17:219–250.

J. Rosselló. 2002. El SV, I: verb i arguments verbals. In Joan Solà (ed.), *Gramàtica del Català Contemporani*, pages 1853–1949. Empúries, Barcelona.

S. Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pages 747–753, Saarbruecken, Germany.

E. Vallduví and E. Engdahl. 1996. The Linguistic Rrealization of Information Packaging. *Linguistics*, 34:459–519.