

Automatic Learning of Syntactic Verb Classes

Laia Mayol and Gemma Boleda and Toni Badia

GLiCom

Dept. of Translation and Philology

Pompeu Fabra University

{laia.mayol,gemma.boleda,toni.badia}@upf.edu

Abstract

This paper describes an experiment devised to group Catalan verbs according to their syntactic behavior. Our goal is to acquire a small number of basic classes with a high level of accuracy, from relatively knowledge-poor resources. This information, expensive and slow to compile by hand, is useful for any NLP task requiring specific lexical information.

The experiment aims at automatically classifying verbs into transitive, intransitive and verbs alternating with a *se*-construction. We use a clustering methodology applied to data extracted from a tagged corpus. Our system achieves an average 0.87 F-score, for a task with a 0.65 baseline. The cluster analysis also provides insight into the relevant features and the notion of prototypicality within a class.

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, specially in the case of verbs, as they have a great influence in 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 lexical properties of words from their behavior in corpora and other resources, using machine learning techniques.

The first works on automatic acquisition of subcategorisation information were not directed at classifying verbs but at compiling every possible subcategorisation frame for each verb. Brent

(1993) used raw corpora to obtain six different frame types; Manning (1993) described a system which could recognize 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 also been used to classify verbs into semantic classes (Schulte im Walde, 2000; Stevenson and Joanis, 2003).

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 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.

In unsupervised techniques, such as clustering, the algorithms do not need any set of pre-classified training instances to compute the solution. Hence, their results are independent from any human classification and depend only on the features and the parameters chosen. It is sensible, therefore, to consider unsupervised techniques to be more empirical than supervised techniques (because the latter do depend on a previous classification).

The paper has the following structure: Section 2 introduces the classification sought; Sec-

tion 3 explains the materials and methodology (data, features and approach) of the experiment, and Section 4 its results; Section 5 lists further work. Finally, Section 6 presents the conclusions of this paper.

2 Classification

Our initial aim was to distinguish between transitive verbs (those subcategorising for an NP or clausal 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. When computing two clusters, transitive verbs were concentrated in a cluster and intransitive and prepositional verbs in the other one, according to expectations. However, and consistently across experimental settings, when computing more than two clusters, the algorithm made divisions of the transitive cluster, and did not separate intransitive from prepositional verbs.

We believe that this is due to the fact that both intransitives and prepositionals cooccur with prepositions and, therefore, they are not different enough to be classified in different clusters. Also, transitive verbs were divided into subclasses because they show a more heterogeneous behavior and its number is much greater than the number of both intransitive and prepositional verbs (see in Section 3.1).

As for the divisions within transitives, they were by no means random. A particular class of verbs tended to be separated from more prototypical transitives: 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 example 1 shows.² We call this class VASE (Verbs Alternating with a *SE*-construction).

- (1) a. La revolució no **beneficia** tothom
the revolution not benefits everyone
‘Revolution doesn’t benefit everyone’

¹*Se* is a morpheme present in the grammar of most Romance languages, which typically absorbs an argument of the verb. There is still debate on whether it absorbs the internal or the external argument. See Bartra (2002) for an overview of its uses in Catalan.

²All examples in the paper are taken from the CTILC corpus (see Section 3.1) and shown literally or in an simplified version.

- b. L’agricultura es **beneficia** del
the agriculture SE benefits of the
conflicte
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 (Rosselló, 2002). In our Gold Standard it corresponds to 20% of the lemmata (opposed to 10% intransitive and 8% prepositional; see Section 3.1). Due to the importance of this alternation, and to the fact that these verbs share properties both with transitive and prepositional verbs (they sometimes bear an NP object, sometimes a prepositional one), we found it advisable to add this class to our targeted classification.

In the light of these experimental results, we redefined the classification and designed a two step procedure. In the first step (Sections 3 and 4), the task was to classify verbs into transitive, intransitive and VASE. Intransitives include both verbs subcategorising prepositional objects and pure intransitives. In the second step, briefly explained in Section 5, we further distinguished between prepositional verbs and pure intransitives.

3 Material and method

3.1 Data: Corpus and Gold Standard

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

The experiments were carried out on 200 verbs, randomly selected among those having more than 50 occurrences in the corpus. 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. The resulting Gold Standard classification is depicted in Table 1.

	#	%
Transitive	129	64.5
VASE	39	19.5
Intransitive	32	16.0

Table 1: Classes for the Gold Standard.

Note that the largest class is by far that of transitive verbs, and that the intransitive class

abbrev.	gloss
1 ObjCl	Cooccurrence with an object clitic.
2 DetOrN	Determiner or noun follows.
3 Passive	Passive construction.
4 Punct	Punctuation marks (stop, colon, etc.) follow.
5 Prep	Preposition follows (except for <i>per</i> ‘by’).
6 Se	Particle <i>se</i> precedes or follows the verb.
7 DetAndN	DetOrN + precedence by an NP element (adj, pron, det or noun).
8 NonAgrN	DetOrN + not agreement in number.
9 NonAgrP	DetOrN + not agreement in person.
10 NonFin	DetOrN + verb in a nonfinite form.

Table 2: Features used for verb classification.

is the smallest one, despite the fact that it includes verbs bearing a prepositional object and verbs with very unfrequent transitive uses (*dormir la migdiada* ‘take a nap’, as transitive use of *dormir* ‘sleep’). Taking this distribution into account, we can establish a baseline for the evaluation: Instead of randomly assigning verbs to classes, we will use a higher baseline, that of assigning all verbs to the most common class, transitive verbs. This results in a 0.65 F-score (more details in Section 4).

3.2 Features

We defined ten features suitable to characterise the targeted classes, along with superficial linguistic cues which allowed us to automatically extract the data by simple frequency counts. Table 2 summarizes the features and the shallow cues, and we describe our hypotheses with respect to the characterisation of the classes in what follows.

The first three features, ObjCl, DetOrN and Passive, are directed towards characterising transitive uses of verbs. We expect transitive verbs to have the highest values for these features, while VASE verbs will have middle values but still higher than intransitive ones, due to the uses of VASE verbs where they occur with an NP object.

Note that, as subjects may appear postverbally in Catalan (especially with unaccusative verbs; see sentence (2)), some intransitive verbs may also have relatively high values for feature DetOrN.

- (2) **Apareixerà** el monstre
 Appear-fut the monster
 ‘The monster will appear’

The following two features, Punct and Prep, are expected to characterise 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.

Feature Se is the only one specifically designed to identify VASE verbs. VASE verbs should have the highest values for this feature and intransitive ones the lowest, since *se* is mostly related to phenomena related to transitivity: reflexivity, passivization, etc.

The last four features, DetAndN, NonAgrN, NonAgrP and NonFin, are aimed specifically at distinguishing transitive verbs from intransitive verbs with a postverbal subject, which is a major problem for our task, as mentioned above and exemplified in sentence (2). The same problem would arise with any other language with a similar syntactic pattern, such as Italian or Spanish. The last features are elaborations on DetOrN designed to detect objects. The restriction that an NP both precedes and follows a verb (feature DetAndN) makes it more likely that an object is present; also, the fact that the NP following the verb does not agree with it in number or person (features NonAgrN and NonAgrP) also point to an object. As for feature NonFin, it exploits the fact that postverbal subjects with infinitives are very rare in Romance languages.

The first six features are represented in terms of raw percentages. Because the last four features are prone to sparse data problems, their values are proportions within the values for DetOrN. The result of the feature extraction is a representation for each verb as in Table 3. We see there e.g. that 9.3% of the occurrences of the verb *contemplar* ‘contemplate’ (transitive) exhibit the feature ObjCl, while *beneficiar* ‘benefit’ (VASE) only presents 3% and *xisclar* ‘scream’ (intransitive) 0%.

Table 4 shows the mean values for each fea-

Lemma	Class	ObjCl	DetOrN	Passive	Punct	Prep
<i>contemplar</i>	Trans.	9.3	52.2	3.4	4.3	15.0
<i>beneficiar</i>	VASE	3.0	20.1	2.5	6.5	32.6
<i>xisclar</i>	Intr.	0	11.7	0	22.0	11.0
		Se	DetAndN	NonAgrN	NonAgrP	NonFin
<i>contemplar</i>	Trans.	5.9	15.1	17.3	13.7	25.4
<i>beneficiar</i>	VASE	37.6	39.2	33.3	3.9	19.0
<i>xisclar</i>	Intr.	0.8	0	0	0	6.6

Table 3: Feature values for verbs *contemplar*, *beneficiar*, and *xisclar*.

ture according to the class.³ Most of the expectations are met: Transitive verbs have the highest values across classes for seven out of the ten features: ObjCl, DetOrN, Passive, DetAndN, NonAgrN, NonAgrP and NonFin. Intransitive verbs have highest values only for Punct and Prep. VASE verbs have intermediate values 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 and may not even be significant, but the patterns are very consistent with our hypotheses.

Feature	Trans.	VASE	Intr.
ObjCl	4.8	<i>4.6</i>	0.5
DetOrN	26.4	<i>16.3</i>	14.1
Passive	6.5	<i>3.1</i>	0.6
Punct	<i>7.1</i>	6.8	10.9
Prep	17.3	<i>31.3</i>	40.2
Se	<i>11.8</i>	33.8	2.6
DetAndN	31.9	<i>27.6</i>	23.0
NonAgrN	28.4	<i>26.5</i>	13.2
NonAgrP	12.4	<i>12.2</i>	3.1
NonFin	54.6	<i>41.7</i>	18.6

Table 4: Mean values for features by class.

3.3 Clustering approach

We used CLUTO⁴ for the experiments. We will report the results obtained with the k -means algorithm. We also tried several of the other algorithms provided with CLUTO (hierarchical and flat, agglomerative and partitional), obtaining quite similar results.

4 Results

With k -means, the number of clusters has to be predetermined. Because our targeted classification consists of three classes, we concentrated

on the three cluster solution and will report results for this partition only. As we see in Table 5, 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. However, as detailed in Table 5, there are also some misclassified verbs, which will be further analysed in Section 4.1. Table 6 shows the mean value for each feature in each cluster.

Cluster	Trans.	VASE	Intr.	Total
0	115	7	5	<i>127</i>
1	9	0	26	<i>35</i>
2	5	32	1	<i>38</i>
Total	<i>129</i>	<i>39</i>	<i>32</i>	<i>200</i>

Table 5: Contingency table.

Feature	0	2	1
ObjCl	5.2	<i>4.0</i>	0.4
DetOrN	26.8	<i>16.2</i>	14.0
Passive	6.7	<i>2.9</i>	1.0
Punct	<i>7.2</i>	6.9	10.2
Prep	15.2	<i>33.4</i>	44.3
Se	<i>10.9</i>	38.9	2.7
DetAndN	31.2	<i>29.2</i>	24.6
NonAgrN	29.6	<i>26.0</i>	10.8
NonAgrP	12.9	<i>10.6</i>	4.3
NonFin	57.6	<i>38.7</i>	14.1

Table 6: Mean values for every feature for clusters 0, 1 and 2

These data fit with the distribution of feature values across classes reported in Table 4, showing that the value distribution of the features defined for each class is consistent with the predictions. For example, verbs which have middle values for features indicating transitivity tend to have a relatively high value for *Se*.

Table 7 shows the evaluation measures as compared to the Gold Standard: Precision, recall and F-score. As for the baseline, recall

³In Tables 4 and 6, the highest mean value appears in bold face, and the middle mean value in italics.

⁴<http://www-users.cs.umn.edu/~karypis/cluto/>.

from Section 3.1 that we use that of considering all verbs to be transitive, the largest class in the Gold Standard. The overall measures are weighted according to the number of verbs in each class, so that they should be read as the probability of correctly classifying a verb, given the distribution of the Gold Standard across classes.

Class	Prec.	Recall	F-score
	Cl. (Bl.)	Cl. (Bl.)	Cl. (Bl.)
Trans.	.91 (.65)	.89 (1)	.90 (.82)
Intr.	.74 (0)	.81 (0)	.78 (0)
VASE	.84 (0)	.82 (0)	.83 (0)
Overall	.87 (.65)	.87 (.65)	.87 (.65)

Table 7: Clustering results (Cl.) compared to baseline (Bl.).

The average F-score is 0.87 a good overall result for a lexical acquisition task, and also compared to the baseline (0.65).

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 features are characteristic of transitives, so that the clustering algorithm has richer information 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. A second experiment we performed was devoted to that distinction (see Section 5).

4.1 Error analysis

Transitive verbs misclassified into cluster 1-Intr.: *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 either are 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 subcategorizes for a different frame. For instance, in the ‘stop’ sense *cessar* subcategorizes for a prepositional phrase, while in the ‘dismiss’ sense it is a plain transitive. We didn’t establish a specific class for this alternation and therefore classified this verb as transitive. As the ‘stop’ sense is far more frequent, the feature values for this verb are closer to those of intransitive verbs and, accordingly, it is classified in cluster 1. This second kind of mis-

take thus points to a richer classification, and the eventual need to encode different frames associated to different senses in case of polysemy. However, this implies a richer lexical representation, which is more difficult to exploit.

Transitive verbs misclassified into cluster 2-VASE: *avorrir* (bore), *coure* (cook), *errar* (err), *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 non-causative 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.

Misclassified intransitive verbs: *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 the verbs. For instance, *esmorzar* and *regalimar* have some transitive uses and *agradar* and *néixer* appear almost exclusively with a postverbal subject.

Misclassified VASE verbs: *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 appear often 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.

To sum up, we have seen that the verbs that have been misclassified are in one way or another not **prototypical** within their class. Intuitively, they should also not be similar to the prototype of the class where they have been wrongly placed. A preliminary analysis of the *z*-scores of the verbs indicate that the intuition is correct for transitive and VASE verbs, but not for intransitive verbs. For two of the clusters, thus, we find that mistakes correspond to distance to the centroid. This suggests that cluster analysis could be used to approach the notion of prototypicality within a class, although further research is needed on this issue.

5 Further work

We are currently testing the system with a 208 million word corpus extracted from the Web (Boleda et al., 2005). With this resource, results are much worse, achieving only a 0.73 F-score (which is however still well beyond the baseline). It is surprising that with on average 12 times the evidence for a verb, results decrease so much; the reason could be the noise contained in such a corpus.

In addition, we have performed another classification experiment which we cannot fully explain due to space constraints. The experiment was aimed at further dividing intransitive verbs into pure intransitives and verbs bearing a prepositional object. The baseline for the task was 0.5 and the upperbound 0.94. Using the experimental setting explained in Section 3 and four features, we achieved an average 0.84 F-score, only 10 points away from the upperbound.

As in the previous experiment, misclassified verbs are verbs whose behavior is closer to the behavior of the verbs of the other class. Most of the misclassified pure intransitives are verbs that very frequently appear with a particular kind of locative adjunct (*conduir per* ‘drive on’, *xocar contra* ‘crash into’). As for misclassified prepositional verbs, they are those which 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’).

6 Conclusions

We have presented a cluster analysis which can be used to classify verbs into basic syntactic classes in Catalan using very simple resources (a corpus with morphological information), and which we believe can be straightforwardly extended to other Romance languages, for which there are typically less available resources than for English.

We classified verbs into transitive, intransitive and verbs alternating with a *se*-construction. We defined ten features with their associated shallow cues, which are linguistically motivated and which our experiments have empirically validated. We achieved a mean F-score of 0.87 for an experiment with a 0.65 baseline, which is a good result for a lexical acquisition task.

We have argued that the defined features and the cluster analysis are also useful to determine

the prototypicality of a verb within a class. Misclassified verbs are those that have some special property (belong to a subclass, present a particular alternation) and, hence, tend to be further from the centroid of the cluster. Therefore, the mistakes of this system are also linguistically motivated.

7 Acknowledgements

Many thanks to Toni Martí, 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 Nadjat Bouayad and Sebastian Padó for revision of previous versions of this paper. This work is supported by the Departament d’Universitats, Recerca i Societat de la Informació (grants 2003FI-00867 and 2001FI-00582).

References

- A. Bartra. 2002. La passiva i les construccions que s’hi relacionen. In J. Solà, editor, *Gramàtica del Català Contemporani*, pages 2111–2179. Empúries, Barcelona.
- G. Boleda, S. Bott, B. Poblete, C. Castillo, M.E. Fuenmayor, T. Badia, and V. Lopez. 2005. Cucweb: A catalan corpus built from the web. In preparation.
- M. Brent. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 7:243–262.
- T. Briscoe and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of ANLP-97*, Washington, USA.
- 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.
- C. Manning. 1993. Automatic acquisition of a large subcategorisation dictionary from corpora. In *Proceedings of the 31st ACL*, pages 235–242.
- 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 J. Solà, editor, *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 COLING-00*, pages 747–753.
- S. Stevenson and E. Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of CoNLL-2003*.
- E. Vallduví and E. Engdahl. 1996. The Linguistic Realization of Information Packaging. *Linguistics*, 34:459–519.