

Acquisition of Semantic Classes for Adjectives from Distributional Evidence

Gemma Boleda

GLiCom

Universitat Pompeu Fabra

La Rambla 30-32

08002 Barcelona

gemma.boleda@upf.edu

Toni Badia

GLiCom

Universitat Pompeu Fabra

La Rambla 30-32

08002 Barcelona

toni.badia@upf.edu

Eloi Batlle

Audiovisual Institute

Universitat Pompeu Fabra

Pg. Circumval.lació 8

08003 Barcelona

eloi@iua.upf.es

Abstract

In this paper, we present a clustering experiment directed at the acquisition of semantic classes for adjectives in Catalan, using only shallow distributional features.

We define a broad-coverage classification for adjectives based on Ontological Semantics. We classify along two parameters (number of arguments and ontological kind of denotation), achieving reliable agreement results among human judges. The clustering procedure achieves a comparable agreement score for one of the parameters, and a little lower for the other.

1 Introduction

The main hypothesis underlying the tasks in Lexical Acquisition is that it is possible to infer lexical properties from distributional evidence, taken as a generalisation of a word's linguistic behaviour in corpora. The need for the automatic acquisition of lexical information arised from the so-called "lexical bottleneck" in NLP systems: no matter whether symbolic or statistical, all systems need more and more lexical information in order to be able to predict a word's behaviour, and this information is very hard and costly to encode manually.

In recent research in the field, the main effort has been to infer semantic classes for verbs, in English (Stevenson et al., 1999) and German (Schulte im Walde and Brew, 2002). In this paper, we concentrate on adjectives, which have received less attention (see though Hatzivassiloglou and McKeown (1993) and Lapata (2001)). Our aim is to establish semantic classes for adjectives in Catalan by means of clustering, using only shallow syntactic evidence. We compare the results with a set of adjectives classified by human judges according to semantic characteristics. Thus, we intend to induce semantic properties from syntactic distribution. We now justify each of the choices: why adjectives, why clustering, and why shallow features.

Adjectives are predicates, equivalent to verbs when appearing in predicative environments. A

broad semantic classification like the one we propose is a first step for characterising their meaning and argument structure. In their modifying function, they are crucial in restricting the referents of NPs. A good characterisation of their semantics can help identify referents in a given (con)text in dialog-based tasks, Question Answering systems, or even advanced Information Extraction tasks.

We believe clustering, an unsupervised technique, to be particularly well suited for our task because there is no well-established classification we can rely on, so that data exploration is advisable for our task. In clustering, objects are grouped together according to their feature value distribution, not to a predefined classification (as is the case when using supervised techniques), so that we achieve a better guarantee that we are learning a structure already present in the data.

Although adjectives are predicates, they have a much more limited distribution than verbs, and do not present long-distance dependencies. Therefore, we expect that shallow distributional features will be enough for our task. One of the purposes of the paper is to test whether this hypothesis is right. This would make adjective classification achievable for languages with less available resources than English, such as Catalan.

The paper is structured as follows: Section 2 introduces the classification we are aiming at and the hypotheses that led to the experiments; Section 3 focuses on the methodology used to produce the classification; in Section 4 we discuss the results obtained so far; finally, Section 5 contains some conclusions and proposals for further work.

2 Classification and Hypothesis

As mentioned above, the semantic classification of adjectives is not settled in theoretical linguistics. Much research in formal semantics has focused on relatively minor classes (see Hamann (1991) for an overview), which causes coverage problems for NLP systems. Standard descriptive grammars do propose broad-coverage classifications (see Picallo

(2002) for Catalan), but these usually mix morphological, syntactic and semantic criteria. We therefore turned to classifications tailored for NLP systems, and defined two parameters largely inspired by Raskin and Nirenburg (1995):

- *unary* or *binary* adjectives, according to whether they have one or two arguments.
- *basic*, *object* or *event* adjectives, according to whether they denote non-decomposable properties, or it can be postulated that they have an *object* or *event* component in their meaning.

This classification was originally devised for systems using an external ontology (so that semantic representations are directly linked to concepts in the ontology), but it is also suitable for broader settings, as we argue in the rest of the Section. We now turn to briefly present the syntax of adjectives in Catalan and discuss the parameters in more detail.

2.1 Syntax

The default function of the adjective in Catalan is that of modifying a noun; the default position is the postnominal one (about 66% of adjective tokens in the corpus used for the experiments modify nouns postnominally). However, some adjectives can appear prenominal, mainly when used non-restrictively (so-called “epithets”; 26% of the tokens occur in prenominal position).

The other main function of the adjective is that of predicate in a copular sentence (6% of the tokens). Other predicative contexts, such as adjunct predicates, are much less frequent (approx. 1% of the adjectives in the corpus).

2.2 Unary vs. binary

Unary adjectives have only one argument, usually corresponding to the modified noun (*a red ball*_{ARG1}) or the subject in a copular sentence (*this ball*_{ARG1} *is red*). Binary adjectives have two arguments, one analogous to ARG1 and another one which usually corresponds to a PP complement (*a teacher*_{ARG1} *jealous of Mary*_{ARG2}, *this teacher*_{ARG1} *is jealous of Mary*_{ARG2}). Thus, unary adjectives denote properties and binary adjectives denote relations.

From a linguistic point of view, we expect binary adjectives to co-occur with postponed prepositions with a significant higher frequency than unary ones. Similarly, because of the heaviness of the PP, we expect them to frequently occur in predicative constructions, that is, after a verb.

The arity is a basic parameter for the semantic characterisation of any predicate. It is use-

ful for low-level tasks such as parsing (e.g. for PP-attachment ambiguity within NPs), but also for tasks oriented to semantics, such as the extraction of relationships between individuals or concepts.

2.3 Basic denotation vs. object component vs. event component

Basic adjectives denote attributes or properties which cannot be decomposed; for instance, *red* or *jealous*. Adjectives which have an event component in their meaning (*event adjectives* for short) denote a state that is directly dependent on an event, be it simultaneous or previous to the state. Examples would be *directed*, *flipping* or *constitutive*. Similarly, *object adjectives* have an embedded object component in their meaning: *pulmonary disease* can be paraphrased as *disease that affects the lungs*, so that *pulmonary* evokes the object *lungs*. Other examples would be *economic* or *agricultural*.¹

We expect object adjectives to have a rigid position, right after a noun (in Catalan). Any other modifiers or complements (PPs, other adjectives, etc.) will occur after the object adjective. This restriction also implies that they will have very low frequencies for predicative positions.

Event adjectives, on the contrary, appear most naturally in predicative environments. This is probably due to the fact that most of them are deverbal and thus inherit part of the verbal argument structure. Thus, they tend to form larger constituents that are mostly placed in predicative position. For the same reason, they will appear in postnominal position when acting as modifiers.

As for basic adjectives, most of them can be used nonrestrictively, so that they will appear both postnominally and prenominal. In addition, there is no restriction keeping them from appearing in predicative constructions. When combined with other kinds of adjectives, mainly object adjectives, they will appear at the periphery (*anàlisi política seriosa*, ‘serious political analysis’).

This parameter can again be used for basic tasks such as POS-tagging: Adjective-noun ambiguity is notoriously the most difficult one to solve, and the ordering restrictions on the classes of adjectives can help to reduce it. However, it is most useful for semantic tasks. For instance, object adjectives can evoke arguments when combined with predicative nouns (*presidential visit* - a president visits X). For projects such as FrameNet (Baker et al., 1998),

¹Note that we do not state that adjectives denote objects or events, but that they imply an object or event in their denotation. This kind of adjectives denotes properties or states, but with an embedded or “shadow” argument (Pustejovsky, 1995), similarly to verbs like *to butter*.

these kinds of relationships could be automatically extracted if information on the class were available. The same applies to event adjectives, this time being predicates (*flipping coin* - a coin flips).

2.4 Morphology vs. syntax

It could seem that the semantic classes established for the second parameter amount to morphological classes: not derived (basic adjectives), denominal (object adjectives), and deverbal (event adjectives). However, although there is indeed a certain correlation between morphological class and semantic class, we claim that morphology is not sufficient for a reliable classification because it is by no means a one-to-one relationship.

There are denominal adjectives which are basic, depending on the suffix (e.g. *-ós* as in *vergonyós*, ‘shy’) and on whether they have developed a different meaning than the etymological one, such as *marginal*, ‘marginal’, which has come to be used as synonymous to ‘rare, outsider-like’. Conversely, some object adjectives are not synchronically denominal, such as *botànic*, ‘botanical’. The same happens with event as opposed to deverbal adjectives: a deverbal adjective such as *amable* (lit. ‘suitable to be loved’, has derived to ‘kind, friendly’) has now a basic meaning (we have not found any non-deverbal adjective to have an event-type denotation).

Our hypothesis, which will be tested on Section 4.3, is that syntax is more reliable than morphology as a basis for semantic classification. The intuition behind this hypothesis is that if a certain suffix forms basic adjectives, they will behave like ordinary basic adjectives; similarly, if a derived adjective has undergone semantic change and as a result has shifted class, it will also behave like an ordinary adjective of the target class.

3 Methodology

We used a 16.5 million word Catalan corpus, semi-automatically morphologically tagged and hand-corrected (Rafel, 1994). The corpus contains modern written samples (1960-1988) from most topics and genres. We selected all adjectives in the corpus with more than 50 occurrences (2283 lemmata), including some gerunds and participles with a predominant modifying function (for more details on the selection criteria, cf. Sanromà (2003)).

In all the experiments, we clustered the whole set of 2283 adjectives, as the set of objects alters the vector space and thus the classification results. We therefore clustered always the same set and chose different subsets of the data in the evaluation and testing phases in order to analyse the results.

tag	gloss	tag	gloss
*cd	clause delimiter	aj	adjective
*dd	def. determiner	av	adverb
*id	indef. det.	cn	common noun
*pe	preposition	co	coordinating elem.
*ve	verb	np	noun phrase
ey	empty		

Table 1: Tags used in the bigram representation. Phrase boundary markers signaled with *.

In the evaluation phase we used a manually classified subset of 100 adjectives (*tuning subset* from now on). Two judges classified them along the two parameters explained in Section 2 and their judgments were merged by one of the authors of the paper. In the testing phase, we used a different subset with 80 adjectives as Gold Standard against which we could compare the clustering results (see Section 3.2 for details on the manual annotation process).

3.1 Feature representation

Although we already had some hypotheses with respect to what features could be relevant, as discussed in Section 2, we wanted to proceed as empirically as possible. Recall also from the Introduction that we wanted to restrict ourselves to shallow distributional features. For both reasons, we modelled the data in terms of blind n-gram distribution and then selected the features.

The lemmata were modelled using pairs of bigrams: in a 4-word window (three to the left and one to the right of the adjective), the first two tags formed a feature and the second two tags another feature. They were encoded separately due to sparse data considerations. This window should be enough for the kind of information we gather, because of the locality of the relationships which most adjectives establish with their arguments (see Section 2).

We subsumed the information in the original morphological tags in order to have the minimal number of categories needed for our task, listed in Table 1.²

In order to further reduce the number of features in a linguistically principled way, we took phrase boundaries into account: All words beyond a POS considered to be a phrase boundary marker (see Table 1) were assigned the tag *empty*.

Examples 1 and 2 show the representation that would be obtained for two imaginary English sen-

²Clause delimiters are punctuation marks other than commas, relative pronouns and subordinating conjunctions. Coordinating elements are commas and coordinating conjunctions. Noun phrases are proper nouns and personal pronouns. Clitic pronouns were tagged as verbs, for they always immediately precede or follow a verb.

tences (target adjective in bold face, word window in italics; negative numbers indicate positions to the left, positive ones positions to the right):

1. He says that the **red** ball is the one on the left.
-3ey-2cd, -1dd+1cn.
2. Hey, this teacher is **jealous** of Mary!
-3ey-2ey, -1ve+1pe.

The representation for sentence 1 states that the first element of the 5-gram (-3; third word to the left of the adjective) is empty (because the second element is a phrase boundary marker), that the second element is a clause delimiter (conjunction *that*), the third one (-1; word preceding the adjective) is a definite determiner, and the fourth one (+1; word following the adjective) is a common noun.

This representation schema produced a total of 240 different feature (bigram) types, 164 of which had a prior probability < 0.001 and were discarded. In order to choose the most adequate features for each of the parameters (that is, features that allowed us to distinguish unary from binary adjectives, on the one hand, and basic from event and from object adjectives, on the other), we checked the distributions of their values in the *tuning subset*. Features were chosen if they had different distributions in the different classes of each parameter and they made linguistic sense. We found that both criteria usually agreed, so that the selected features are consistent with the predictions made in Section 2, as will be discussed in Section 4. An alternative, more objective selection method would be to perform ANOVA, which we plan to test in the near future.

3.2 Gold Standard

Recall that we could not use any previously well-established classification. We therefore built our own Gold Standard, as has been mentioned at the beginning of this section.

The 80 lemmata were independently annotated by three human judges (PhD students in Computational Linguistics, two of which had done research on adjectives), who had to classify each adjective as either unary or binary, on the one hand, and either basic, event or object-denoting, on the other. They received instructions which referred *only to semantic characteristics*, not to the expected syntactic behaviour. For example, “check whether the state denoted by the adjective is necessarily related to a previous or simultaneous event”. In addition, they were provided with (the same randomly chosen) 18 examples from the corpus for each of the adjectives to be tagged.

The judges were allowed to assign a lemma to a second category in case of polysemy (e.g. *econdmic* has an object meaning, ‘economic’, and a basic one, ‘cheap’, less frequent in the corpus). However, the agreement scores for polysemy judgments were not significant at all. We cannot perform any analysis on the clustering results with respect to polysemy until reliable scores are obtained.³ We therefore ignored polysemy judgements and considered only the main (first) class assigned by each judge for all subsequent analyses.

The three classifications were again merged by one of the authors of the paper into a single Gold Standard set (GS from now on). The agreement of the judges amongst themselves and with the GS with respect to the main class of each adjective can be found in Tables 2 and 3.

	J1		J2		J3	
	%agr	κ	%agr	κ	%agr	κ
J2	0.88	0.59				
J3	0.98	0.91	0.90	0.67		
GS	0.97	0.89	0.90	0.65	0.98	0.90

Table 2: Agreement for the unary/binary parameter: inter-judge (J1, J2, J3), and with GS

	J1		J2		J3	
	%agr	κ	%agr	κ	%agr	κ
J2	0.83	0.74				
J3	0.88	0.80	0.80	0.68		
GS	0.93	0.89	0.83	0.74	0.92	0.87

Table 3: Agreement for the basic/event/object parameter: inter-judge (J1, J2, J3), and with GS

As can be seen, the agreement among judges is remarkably high for a lexical semantics task: All but one values of the kappa statistics are above 0.6 (+/-0.13 for a 95% confidence interval). The lowest agreement scores are those of J2, the only judge who had not done research on adjectives. This suggests that this judge is an outsider and that the level of expertise needed for humans to perform this kind of classification is quite high. However, there are too few data for this suspicion to be statistically testable.

Landis and Koch (1977) consider values $\kappa > 0.61$ to indicate a substantial agreement, whereas

³The low agreement is probably the result of both the fuzziness of the limits between polysemy and vagueness for adjectives, and the way the instructions were written, as they induced judges to make hard choices and did not state clearly enough the conditions under which an item could be classified in more than one class.

Carletta (1996) says that $0.67 < \kappa < 0.8$ allows just “tentative conclusions to be drawn”. Merlo and Stevenson (2001) report inter-judge κ values of 0.53 to 0.66 for a task we consider to be comparable to ours, that of classifying verbs into unergative, unaccusative and object-drop, and argue that Carletta’s “is too stringent a scale for our task, which is qualitatively quite different from content analysis” (Merlo and Stevenson, 2001, 396).

The results reported in Tables 2 and 3 are significantly higher than those of Merlo and Stevenson (2001). Although they are still not all above 0.8, as would be desirable according to Carletta, we consider them to be strong enough to back up both the classification and the feasibility of the task by humans. Thus, we will use GS as the reference for clustering analysis.

4 Results

The experiments were performed using CLUTO,⁴ a free clustering toolkit. We tested the several clustering approaches available in the tool: two hierarchical and one flat algorithm, one of them agglomerative and the other two partitional, with several criterion functions, always using the cosine distance measure. Two different combinations of features and feature normalisations were tested for each parameter. The best result was obtained with the k-means algorithm and the parameters listed in Table 4. However, the results were quite robust through all parametrisations.⁵

	un, bin	bas, ev, obj
number of clusters	2	3
number of features	10	32
feature normalisation	none	$P(F_j A_i)/P(F_j)$

Table 4: Parameters for the clustering solutions.

4.1 Unary vs. binary

Figure 1 depicts the clustering solution for the unary/binary parameter.

The agreement between GS and this clustering solution resulted in 0.97% and $\kappa=0.87$ (κ ranging from 0.67 to 0.89 with human judges), thus fully comparable to the interjudge agreement.

As can be seen in Figure 1, all binary adjectives are together in cluster 1, while most unary ones are

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

⁵The feature normalisation for the basic/event/object parameter was as follows: for each adjective i and feature j , the raw percentage $P(F_j|A_i)$ was divided by the prior probability of the feature $P(F_j)$, so that the distance from the expected percentage, rather than the percentage value as such, was obtained.

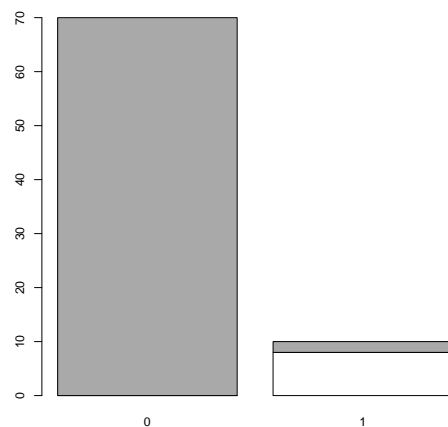


Figure 1: Clustering solution A: clusters (columns) vs. unary (gray) and binary (white) adjectives.

in cluster 0 (only 2 unary adjectives were misclassified as binary). The clustering clearly recognizes a majority of objects bearing no complement and a minority having a regular complement. This parameter, then, is quite easy and reliable to obtain.

Indeed, the most relevant features for each cluster matched very closely the hypotheses discussed in Section 2. They are depicted in Table 5.

cl	high values	low values
0	-1cn+1co, -1cn+1cd	-1aj+1pe, -1ve+1pe
1	-1ve+1pe, -1co+1pe	-1cn+1aj

Table 5: Unary/binary: most relevant features (represented as in examples 1 and 2).

Objects in cluster 1, corresponding to binary adjectives, have high values for most of the features containing a preposition after the adjective (observe +1pe, ‘preposition to the right’). Objects in cluster 0 (unary adjectives), symmetrically, have low values for these features, and high values for the default adjective position in Catalan (directly postnominal: -1cn). The behaviour of the objects in cluster 0 (the biggest cluster by far) presents more cohesion than that of the objects in cluster 1, which have a medium mean value for most features. That is, binary adjectives do not have low values in those features that characterize unary ones, but still significantly lower.

4.2 Basic vs. event vs. object

Figure 2 depicts the clustering solution for the basic/event/object parameter.

The agreement between the GS and the clustering solution was much lower than for the unary/binary

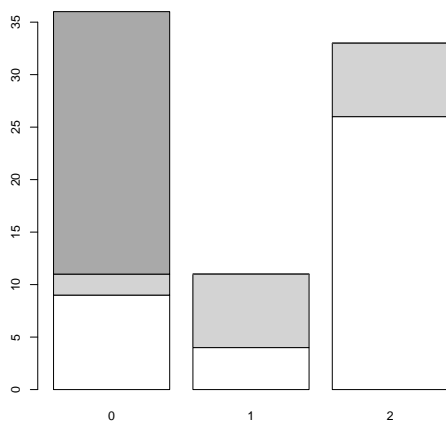


Figure 2: Clustering solution B: clusters (columns) vs. basic (white), event (light gray) and object (dark gray) adjectives.

parameter: 0.73% and $\kappa=0.56$ (+/-0.14 at 95% c.i.; κ ranging from 0.51 to 0.57 with human judges). Our diagnosis is that this is due to the lack of syntactic homogeneity of the event-adjective class, which might be due to a wrong characterisation of the class.

As can be seen in Figure 2, while object adjectives are all in cluster 0 and basic adjectives are concentrated in cluster 2, event adjectives are scattered through clusters 1 and 2. In fact, cluster 1 contains seven out of the eight binary adjectives in GS, and only four unary ones. It seems, then, that what is being spotted in cluster 1 are again binary, rather than event, adjectives. If we look at the morphological type, it turns out that six out of seven event adjectives in cluster 1 (against three out of seven in cluster 2) are participles. A tentative conclusion we can draw is that participles and other kinds of deverbal adjectives do not behave alike; moreover, it seems that other kinds of deverbal adjectives behave quite similarly to basic adjectives.

It should be remarked, however, that although event adjectives do not form a homogeneous class with respect to the features used, basic and object adjectives are quite clearly distinguished from each other in the clustering solution.

As for the features that were most relevant for each cluster, listed in Table 6, they confirm the analysis just made and again match the hypotheses discussed in Section 2.

Lemmata in cluster 0 (object adjectives) have high values for the expected “rigid” position, right after the noun (-1cn) and before any other adjective

cl	high values	low values
0	-1cn+1aj , -1cn+1ve	-1ve+1pe , -1ve+1dd
1	-1ve+1pe , -1cd+1pe	-1cn+1aj , -1co+1cn
2	-1co+1cd , -1co+1co	-1aj+1pe , -1cn+1aj

Table 6: Basic/event/object: most relevant features (represented as in examples 1 and 2 above).

(+1aj). They are further characterised by not occurring as predicates (low value for -1ve). As for objects in cluster 1, their features are very similar to the binary cluster 1 above. Finally, cluster 2 (basic adjectives) presents the predicted flexibility: its adjectives occur in coordinating constructions (-1co, +1co) and appear further from the head noun than other adjectives (low value for -1cn+1aj).

4.3 What about morphology?

One of the hypotheses we wanted to test, as stated in Section 2.4, is that syntactic information is more reliable than morphological information in order to establish semantic classes for adjectives. We therefore expect agreement between the clustering solution and GS to be higher than the agreement with a classification based on morphological class. From the manual annotation in Sanromà (2003), we mapped the classes as in Table 7, following the discussion in Section 2.⁶

morph	sem
not derived	basic
denominal	object
deverbal	event
participle	event

Table 7: Mapping from morphology to semantics.

The agreement between this classification and the GS was 0.65% and $\kappa=0.49$, much lower than the agreement between clustering and GS reported above (0.73% and κ 0.56).

Actually, 13 out of 35 denominal adjectives, 7 out of 13 deverbal adjectives and 5 out of 15 participles were considered to be basic in the GS. Most of these mismatches are caused by changes in meaning (e.g. *mecànic*, ‘mechanical’ does not only mean ‘related to mechanics’, but ‘monotone’). The morphological mapping works best for nonderived adjectives: 14 out of 16 were basic in denotation (the remaining two were classified as object). Thus, our hypothesis seems to be backed up by the data available.

⁶Note that this test cannot be performed for the unary/binary parameter, for there is no clear hypothesis with respect to the morphology-semantics mapping.

5 Conclusions and future work

In this paper we have pursued a line of research that seeks to induce semantic classes for adjectives from distributional evidence. Our current results indicate that it is possible, at least for Catalan. We believe that the approach could be straightforwardly extended to other Indo-European languages, such as Spanish, German or English.

The resulting clusters largely correspond to the targeted classes in both parameters: unary vs. binary on the one hand, and basic-property vs. event-component vs. object-component on the other. This is a remarkable result considering (a) that the human judges based their decisions on semantic criteria, whereas the features used corresponded to shallow distributional evidence, and (b) that we used an unsupervised technique. We have shown that for a part of speech with a limited syntactic distribution such as adjectives, this kind of information is enough to achieve a broad semantic classification.

Our results also indicate that a semantic classification based on syntactic distribution is superior to one based on morphological class, mostly due to cases where the adjective has undergone diachronic change in meaning.

However, there is a class that is not well identified: event adjectives. The clustering only identifies those that are binary, thus simply overlapping with the first parameter. The remaining event adjectives seem to behave like basic ones.

Therefore, the first task in future work will be to review the definition and characterisation of this class. Also, as the present analysis is based on a small sample of manually annotated adjectives, we intend to obtain a larger Gold Standard, in order to establish statistically more reliable results. This will also allow further analysis of the data, e.g. to check to what extent errors in the clustering results correspond to disagreement between human judges; or how far from the centroid are objects for which judges disagree. Further experiments with alternative modelling strategies and clustering algorithms should be also performed, so that a global analysis of the approach can be made.

We would also like to investigate what are the limits of adjective classification using only shallow distributional features, and what kinds of information would be adequate to enrich the modelling. Last but not least, we have to work on the definition of polysemy within our task, so that we can achieve significant agreement scores among judges and integrate this parameter in the experiment.

Acknowledgements

Many thanks to the many people who have manually annotated data at some stage of the research: Àngel Gil, Laia Mayol, Martí Quixal, Roser Sanromà, Clara Soler. Also thanks to Nadjet Bouayad, Katrin Erk and 5 anonymous reviewers for revision and criticism of previous versions of the paper. Special thanks are due to Roser Sanromà for providing us with an electronic version of her manual morphological classification (Sanromà, 2003), and to the Institut d'Estudis Catalans for lending us the research corpus. This work is supported by the Departament d'Universitats, Recerca i Societat de la Informació (grant 2001FI-00582).

References

- C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, Montreal, Canada.
- J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.
- C. Hamann. 1991. Adjectivsemantik/Adjectival Semantics. In von Stechow and Wunderlich, editors, *Semantics. An International Handbook of Contemporary Research*, pages 657–673. De Gruyter, Berlin/NY.
- V. Hatzivassiloglou and K. R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st ACL*, pages 172–182.
- J. R. Landis and G. C. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*.
- M. Lapata. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of the NAACL*, pages 63–70, Pittsburgh.
- P. Merlo and S. Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.
- C. Picallo. 2002. L'adjectiu i el sintagma adjectival. In Joan Solà, editor, *Gramàtica del català contemporani*, pages 1643–1688. Empúries, Barcelona.
- J. Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge.
- J. Rafel. 1994. Un corpus general de referència de la llengua catalana. *Caplletra*, 17:219–250.
- V. Raskin and S. Nirenburg. 1995. Lexical semantics of adjectives: A microtheory of adjectival meaning. Technical report, New Mexico State University.
- R. Sanromà. 2003. Aspectes morfològics i sintàctics dels adjectius en català. Master's thesis, Universitat Pompeu Fabra.
- S. Schulte im Walde and C. Brew. 2002. Inducing German semantic verb classes from purely syntactic subcategorisation information. In *Proceedings of the 40th ACL*, pages 223–230.
- S. Stevenson, P. Merlo, N. Kariaeva, and K. Whitehouse. 1999. Supervised learning of lexical semantic verb classes using frequency distributions. In *Proceedings of SigLex99: Standardizing Lexical Resources*, College Park, Maryland.