Automatic acquisition of semantic classes for adjectives

Ph. D. thesis by

Gemma Boleda Torrent

Under the supervision of

Toni Badia Cardús and Sabine Schulte im Walde

Pompeu Fabra University

Barcelona, December 2006

Dipòsit legal: B.53873-2007
ISBN: 978-84-691-1206-9

Abstract

This thesis concerns the automatic acquisition of semantic classes for adjectives. Our work builds on two hypotheses: first, that some aspects of the semantics of adjectives are not totally unpredictable, but correspond to a set of denotational types (semantic classes). Therefore, adjectives can be grouped together according to their semantic class. Second, that the semantic class of an adjective can be traced in more than one linguistic level. In particular, the morphology-semantics and syntax-semantics interfaces are explored for clues that lead to the acquisition of the targeted semantic classes.

Since we could not rely on a previously established classification, a major effort is devoted to defining an adequate classification. The classification proposal is reached through an iterative methodology. By combining deductive and inductive approaches, we evolve from an initial classification based on literature review to a final classification proposal that takes advantage of the insight gained through a set of experiments.

Each iteration consists of three steps: (a) a classification proposal is made; (b) a number of classification experiments, involving human subjects and machine learning techniques, are carried out; (c) through the analysis of the experiments, advantages and drawbacks of the classification proposal are identified.

We present a total of three iterations. The first one starts with a classification based on literature review. A Gold Standard is built and tested against results obtained through a series of unsupervised experiments. The analysis suggests a refinement of the classification. The second iteration uses the refined classification and validates it through a new annotation task performed by human subjects and a further set of unsupervised experiments.

The third iteration incorporates three significant modifications: first, a large-scale human annotation task with 322 human subjects is carried out. For this task, the estimated agreement (K 0.31 to 0.45) is very low for academic standards. Thus, the achievement of reliable linguistic data is revealed as a major bottleneck. Second, the architecture used for the automatic classification of adjectives is modified so that it allows for the acquisition of multiple classes, so as to account for polysemy. The best results obtained, 84% accuracy, improve upon the baseline by a raw 33%. Third, a systematic comparison between different levels of linguistic description is performed to assess their role in the acquistion task at hand.

Acknowledgements

I am grateful to my supervisors, Toni Badia and Sabine Schulte im Walde, for their support and advice during the development of this thesis. Toni has wisely guided me through the whole process, and, even in his busiest times, has always been there when I needed him. His insightful comments have provided keys to analysis that I would not have found on my own. Sabine has provided valuable advice in methodological and technical issues and has given me a lot of emotional support. She has always had a global vision of this piece of research and its value that has helped me make sense of it.

Louise McNally is a model to me, as a researcher but also as a person. I am grateful to her for many discussions about semantics, and for leading me into fruitful joint research. I am also thankful to Stefan Evert for many discussions about several aspects of my work, and particularly for thorough discussions and many ideas regarding the assessment of interrater agreement. The first piece of research I was involved in was a joint paper with Laura Alonso. Many of her ideas, such as using clustering for the classification of adjectives, shaped the initial steps of this thesis.

During the PhD, I was on two occasions a visiting researcher at the Department of Computational Linguistics and Phonetics (CoLi) in Saarland University. I am very grateful to Manfred Pinkal and the SALSA team for letting me collaborate in their project. In particular, I learned a lot from Sebastian Padó and Katrin Erk. Alissa Melinger provided me with very useful advice for the design of the web experiment reported in this thesis. The work reported has also benefited from discussions with Eloi Batlle, Nedjet Bouayad, Gretel De Cuyper, Teresa Espinal, Adam Kilgarrif, Alexander Koller, Maite Melero, Oana Postolache, and Suzanne Stevenson. I also received helpful feedback when presenting my work at several seminars and conferences.

I am grateful to the GLiCom members for providing a nice working atmosphere and for help in many aspects of the PhD. Eva Bofias introduced me into Perl –her 4-page manual was all I needed for my scripts for many months– and into LaTeX. Stefan Bott has helped me many times with programming difficulties. Àngel Gil, Laia Mayol, and Martí Quixal have annotated data and have entered into useful discussion about the difficulties involved in the semantic classification of adjectives.

Outside GLiCom, the "Barcelona adjective working group" members, Roser Sanromà and Clara Soler, have also participated in annotation. I am specially indebted to Roser for lending me the manually annotated database of adjectives she spent so many hours developing, and for spending three afternoons as part of the expert committee annotating the last Gold Standard. I also want to express my gratitude to the (over 600!) people who took part in the web experiment.

I am infinitely grateful for the existence of much freely available software that has made producing, analysing, and reporting the results of this thesis much easier –or even possible. Some of the tools I find most useful are CLUTO, the IMS Corpus WorkBench, R, TeXnicCenter, Weka, and, of course, Perl.

The research reported here has also received the support of some institutions. Special thanks are due to the Institut d'Estudis Catalans for lending the research corpus to the Pompeu Fabra Uni-

versity. The financial support of the Pompeu Fabra University and its Translation and Philology Department, the Generalitat de Catalunya and the Fundación Caja Madrid is gratefully acknowledged. The administrative staff of the Translation and Philology Department, and in particular Susi Bolós, have helped me solve so many burocratic issues over the last six years that it is impossible to list them here.

On a personal note, I thank my friends and family for their love and support. The *Associació 10 de cada 10* has hosted very entertaining Wednesday evenings for over a decade now. My parents have helped me in my PhD not only as most parents do, by being enormously proud of me. My father taught me how to program when I still was a Philology student, and has even helped in implementation work for the PhD. My mother's experience as a researcher has made it easier to make my way through the "research way of life", and she has helped me overcome many a bottleneck in the development of the thesis.

I am very lucky to have met Martí, my husband. He has had to suffer endless explanations and doubts about almost every aspect of the thesis, and has always provided constructive criticism, even late at night when he was hardly awake. And finally, my daughter Bruna is the light of my life. She has also contributed to this thesis, by regularly (sometimes, even irregularly) switching my mind off work.

Contents

List of figures

List of tables

List of acronyms and other abbreviations

Symbols and conventions:

#	mark of semantically anomalous expression
??	mark of unacceptable expression
*	mark of ungrammaticality
A $\models$ B	A logically entails B
A $\not\models$ B	A does not logically entail B

Abbreviations in bibliographical cites:

ch.	chapter
ex.	example
ff.	and following
p.	page
pp.	pages

Abbreviations for semantic classes of adjectives:

B	basic
BE	polysemous between basic and event-related
BO	polysemous between basic and object-related
E	event-related
I	intensional
IQ	polysemous between intensional and qualitative
O	object-related
Q	qualitative
QR	polysemous between qualitative and relational
R	relational

Abbreviations for linguistic concepts:

1ps	first person singular
3ps	third person singular
adj.	adjective
PAUX	past auxiliary verb
det.	determiner
conj.	conjunction
cop.	copular
compl.	complement
mod.	modifier
Morph.	morphological
N	denominal derivational type
NEG	negative particle
NP	noun phrase
O	not derived derivational type
P	participial derivational type

polys.	polysemous classes
POS	part of speech
PP	prepositional phrase
prep.	preposition
pred	predicate
punct.	punctuation
REFL	reflexive pronoun
V	deverbal derivational type

Other abbreviations and acronyms:

10x10 cv	10 run, 10-fold cross validation
adjac.	adjacent
agr.	agreement
CatCG	Catalan Constraint Grammar
CL	Computational Linguistics
classif.	classification
conf. int.	confidence interval
CTILC	*Corpus Textual Informatitzat de la Llengua Catalana* (Computerised Textual Corpus for the Catalan Language)
def.	definition
e.g.	Latin exempli gratia (“for instance”)
est.	estimate
Exp.	experiment
GLiCom	*Grup de Lingüística Computacional* (Computational Linguistics Group at the Pompeu Fabra University)
GS	Gold Standard
i.e.	Latin id est (“that is”)
M	mean
main cl.	main classification
NLP	Natural Language Processing
SCF	subcategorisation frame
SD	standard deviation
SE	standard error
Trans.	translation
vs.	Latin versus (“as opposed to”)

Chapter 1
Introduction

This thesis concerns the automatic acquisition of semantic classes for adjectives. Our goal is to acquire lexical semantic information, that is, semantic information that is specific to a word or groups of words. The thesis fits into the Lexical Acquisition field within Computational Linguistics, a field that has received much interest since the beginning of the nineties.

The main hypothesis underlying the approaches developed 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" (Zernik, 1991) 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.

The information we want to acquire regards the semantic class of adjectives. Our research it builds on two hypotheses: first, that some aspects of the semantics of adjectives are not totally unpredictable, but correspond to a set of denotational types. Therefore, adjectives can be grouped together according to their semantic type (class), a parameter that we will specify more concretely in Chapter 3. Second, that the semantic class of an adjective can be traced in more than one linguistic level. In particular, the morphology-semantics and syntax-semantics interfaces will be explored for clues that lead to the acquisition of the targeted semantic classes.

This chapter offers motivations for the present line of research, summarises the approach taken to tackle it, and highlights its main contributions. It also offers an overview of the remaining chapters.

1.1 Motivation

The initial motivation for our research arose in developing an NLP system for Catalan within the GLiCom group. The core of this system, named CatCG, consists of a set of manually written grammars that assign part of speech and syntactic information to Catalan texts (see Section 2.2.1). The system does not necessarily yield completely disambiguated output, as most statistical taggers and parsers do.

In an advanced development stage of the tool, we observed that a high proportion of the remaining ambiguity in part of speech tagging involves adjectives; an estimate obtained on the corpus used for this thesis (see Section 2.1.1) is 55% of the remaining ambiguity. Much of the ambiguity involves adjective and noun readings of the same word, which get worsened when ambiguous words cooccur. An example is *general francès*, with 4 possible readings corresponding to different part of speech assignments listed under (1.1).

(1.1) a. general$_{adj}$ francès$_{adj}$
'general$_{adj}$ French$_{adj}$ one'

b. general$_{noun}$ francès$_{adj}$
‘French$_{adj}$ general$_{noun}$’

c. general$_{adj}$ francès$_{noun}$
‘#general$_{adj}$ French/Frenchman$_{noun}$’ or

d. general$_{noun}$ francès$_{noun}$
‘#*French/Frenchman$_{noun}$ general$_{noun}$’

The most plausible reading is ‘French general’, as in (1.1b), but ‘general French one’ (as in *el problema general francès,* ‘the general French problem’), as in (1.1b) is also a possible reading. The other variants would be really weird. To solve the ambiguity at least partially, given the data in (1.1), *francès* could be disambiguated to adjective using a general rule that prefers postnominal modification for adjectives. As we will see in Chapter 3, postnominal modification is the default in Catalan.

However, some very frequent adjectives prefer prenominal modification, as in examples in (1.2).

(1.2) a. petit$_{adj}$ animal$_{noun}$
‘small animal’ (and not: ‘#brute child’)

b. sol$_{adj}$ inspector$_{noun}$
‘only inspector’ (and not: ‘#inspecting sun’)

c. grans$_{adj}$ secrets$_{noun}$
‘big secrets’ (and not: ‘#secret pimples/grains’)

Moreover, the noun-noun construction that we discard for (1.1d) is less frequent in Catalan than in English, but also occurs, as witnessed under (1.3). It cannot be uniformly discarded.

(1.3) a. la paraula corder
‘the word corder’

b. la part nord
‘the northern part’

c. un penyalar color de plom
‘a plumb-colour mountain’.

Intuitively, adjectives such as *petit*, *sol*, and *gran* are semantically different from *francès*, and *francès* shares similarities with adjectives such as *italià* (‘Italian’) or even *capitalista* (‘capitalistic’). However, in most computational dictionaries (including the one used for CatCG), all these words are assigned the same tag, *adjective*. Any distinctions have to be made either on a lemma basis (overseeing the common properties of groups of lemmata), or on a tag basis (thus losing discriminative power).

These difficulties arose within a symbolic processing tool; however, the difficulties concern statistical systems as well, for the information available for statistical systems is also either the general tag or the individual lemma or word form. Indeed, the noun/adjective ambiguity, as well as verb/adjective ambiguity for participles, have long been recognised as the most problematic for humans as well as for NLP tools in languages like English and German (Marcus et al., 1993; Jurafsky and Martin, 2000; Brants, 2000).

The observation of general tendencies in syntactic behaviour of adjectives that correspond to broad semantic properties led us to the task faced in this thesis, namely, to pursue a semantic classification of adjectives. Identifying the class of a particular adjective could serve as an initial schema for its lexical semantic representation, that could be further developed using manual or automatic means. As we will see, the semantic classes correspond to broad sense representations, so that they can be exploited for word sense disambiguation tasks.

Adjectives play an important role in sentential semantics. They are crucial in determining reference of NPs. If in a particular context several students are around, the NP (1.4) can be uttered to point to a particular one.

(1.4) l'estudiant sord
‘the deaf student’

Conversely, they help establish properties of entities. If we hear a sentence like (1.5), we can quite safely infer that the (invented) object referred to by *maimai* is an edible physical object.

(1.5) this maimai is round and sweet

A semantic classification like the one we propose in this thesis is a first step in characterising their meaning. A good characterisation of adjective semantics can help identify referents in a given (con)text in dialog-based tasks or Question Answering systems. In addition, it can help in Information Extraction tasks, through attribute extraction. Semantic classes can also be used for basic tasks such as POS-tagging, to solve difficulties such as the ones exemplified in this Introduction.

The semantic classification is established through an iterative process in which a particular classification is proposed, manual annotation and machine learning experiments are performed, and the analysis of both tasks provides feedback to the proposal, indicating flaws or difficulties in several aspects of the classification.

1.2 Goals of the thesis

The goals of this thesis are the following:

- **To define a broad, consistent, and balanced semantic classification for adjectives** The definition of a semantic classification for adjectives is a controversial matter from a theoretical point of view. We define a classification based on literature review and empirical exploration, and revise it according to the results obtained with machine learning experiments.

- **To test the feasibility of the classification task by humans** Because no previously established Gold Standard existed that we could rely on for the machine learning experiments, a major effort has involved the establishment of a reliably labeled set of adjectives. To achieve this goal, proper methodologies have to be developed for the following tasks:

  - to gather human judgments with respect to the semantic class of adjectives
  - to assess agreement among judges for the task, and to analyse sources of disagreement.

- **To test the feasibility of the classification task by computers** The thesis aims at automating the task of semantically classifying adjectives. The task is automated through the use of unsupervised and supervised machine learning techniques, using several pieces of information that we encode in the form of features. Within this goal, two subgoals can be distinguished:

  - to develop an adequate architecture for the acquisition of multiple classes, so as to account for polysemy.
  - to test the strengths and weaknesses of several linguistic levels of description (morphology, syntax, semantics) for the task at hand.

1.3 Structure of the thesis

The thesis is divided into seven chapters. The first chapter is this Introduction, and the remaining chapters contain the information summarised in what follows.

Chapter 2 presents the tools and the data sources used in this thesis.

Chapter 3 discusses several theoretical approaches to the lexical semantics of adjectives, and their relationship to the goals and proposals of this thesis. It focuses on descriptive grammar, formal semantics, and ontological semantics, as well as the treatment of adjective semantics within NLP resources such as WordNet. It also offers an overview of previous work on Lexical Acquisition for adjectives. The first classification is proposed in this chapter.

Chapter 4 explains the methodology and results of three manual annotation experiments, focusing on a large-scale web experiment designed for the last Gold Standard. It provides a thorough discussion of agreement measures and proposes an approach to assess agreement for the data gathered through the web experiment.

Chapter 5 concerns two unsupervised experiments directed at refining and validating the classification proposal. In these experiments, over 2000 adjectives are clustered and results are analysed through coffmparison with the Gold Standard sets and exploration of the characteristics of the obtained clusters. As a result of these experiments, the classification proposed in Chapter 3 is altered.

Chapter 6 reports on a series of supervised experiments that allow for the acquisition of multiple classes per adjective, thus accounting for polysemy. The semantic classification task is viewed as a multi-label classification task, and the classification is performed in two steps: first, binary decision on each of the classes. Second, merging of the individual decisions to achieve a full classification.

Chapter 7 finishes the discussion with some conclusions and directions for future research.

For clarity purposes, a list of of acronyms and other abbreviations used in this document is provided in page *xi*.

1.4 Contributions

The main contributions in this thesis are listed below.

- A semantic classification proposal for adjectives in Catalan that has broad coverage, is based on a unique parameter, and can be traced at different levels of linguistic description. The classification is valid for Romance languages in general. The classes proposed correspond to coarse sense distinctions for particular adjectives.
- Three sets of manually annotated adjectives, totalling 491 lemmata.
- A method to ellicit semantic classes from naive subjects, defining a web experiment so as to distribute the annotation load and gather large amounts of data (Chapter 4). In the actual experiment, we have gathered data from 322 subjects regarding the semantic class of 210 adjectives (32 to 59 responses per lemma, depending on the adjective).
- A method to assess the degree of agreement among the subjects with respect to the task defined in the web experiment (Chapter 4). Three contributions are made within this method:
  - A weighting scheme that is based on an explicit model of the task. The weighting scheme accounts for partial agreement with respect to polysemy judgments, and can be generalised to other tasks involving polysemy.
  - A sampling approach that allows the computation of confidence intervals for the mean agreement values with different measures. This approach is appropriate for experiments in which a large number of subjects participate.
  - The use of entropy as a measure of intra-item agreement. This measure represents the controversy of a particular item, and has been used to identify kinds of adjectives that are particularly difficult.
- The use of unsupervised techniques (clustering) to re-define the classification proposal (Chapter 5). Although unsupervised techniques have been used in related work in Lexical Acquisition (see for instance Schulte im Walde (2006)), the results are typically used to assess, not to modify the targeted classification.
- An architecture to acquire semantic classes including polysemous classes (Chapter 6). The task has been viewed as a multi-label classification task, and the classification is achieved in two steps: first, a binary decision on each class. Then, a merging of the decisions. The architecture can be adapted to other tasks involving polysemy.
- A detailed study of the performance of different levels of linguistic description with respect to semantic classification, in the setting of a supervised machine learning experiment.
- Generally, the use of the manual annotation tasks, the feature analysis, and the machine learning experiments as tools for linguistic research.

All the data gathered is freely available to the research community (upon request to the author).

Parts of the material presented in this thesis have been published or are currently under submission for publication, as will be signaled at the relevant points in the discussion.

Chapter 2
Tools and resources

This chapter summarises the most relevant characteristics of the resources and tools used for the development of the PhD: the data sources (corpus and adjective database) are described in Section 2.1 and the tools used for various purposes are described in Section 2.2.

2.1 Data

2.1.1 Corpus

All experiments are based on a 14.5 million word fragment of the CTILC corpus[1] (*Corpus Informatitzat de la Llengua Catalana*; Rafel (1994)), developed at the Institut d'Estudis Catalans. The corpus has been semi-automatically annotated and hand-corrected, providing lemma and morphological information (part of speech and inflectional features with an EAGLES-like tagset). Structural information is only provided at the text and paragraph level (no sentence delimitation).

2.1.1.1 Distribution of texts across genre and topic

The whole CTILC corpus contains over 50 million words from over 3,300 texts written between 1833 and 1988. It is a traditional corpus, with pre-defined criteria with respect to its content, defined in terms of genre (for literary texts: 30.6% of the corpus) and topic (for non-literary texts: 69.4% of the corpus). Within literary texts, the genre distribution is as shown in Table 2.1. Table 2.2 lists the distribution of non-literary texts across topics (Rafel, 1994).

genre	**% texts**	**% words**
Essay	12,3	13
Narrative	29,9	60
Poetry	27,6	11
Theater	30,1	16

Table 2.1: *Distribution of literary texts in CTILC across genres.*

The fragment we use in the PhD is the set of most modern texts (1960-1988). Unfortunately, we cannot access their metadata, so we cannot know the distribution of genres or topics within the fragment. Assuming that it is quite similar to the overall distribution, we can foresee some of the characteristics of these data that will affect the whole acquisition process.

First, there are no texts corresponding to transcriptions of oral discourse. The closest to that kind of discourse we get are fragments of theater, but they are only a small percentage of the

[1] Punctuation marks are ommited from the word count. With punctuation mark, the count is 16.5 million words.

topic	% texts	% words
Philosophy	4	6
Religion and Theology	8,2	10,2
Social Science	15,6	19,1
Press	18,6	12,3
Natural Science	6	7,6
Applied Science	13,1	15,3
Art, Leisure, Sports	9,3	9,6
Language and Literature	9,3	7,6
History and Geography	9,8	12
Correspondence	6,1	0,5

Table 2.2: *Distribution of non-literary texts in CTILC across topics.*

corpus (about 5% of the overall volume), and literary oral discourse of course has little to do with real oral discourse.

The linguistic behaviour of adjectives in formal written text is presumably quite different from the behaviour in spontaneous speech[2] The differences possibly affect the choice of features when carrying out lexical acquisition experiments, and also the relative usefulness of each of them. Unfortunately, we do not have access to a sufficiently large oral corpus of Catalan to be able to check these differences or assess their impact in our task.

The lexical choice is also affected by the kind of text the corpus is based on: certain adjectives are more prototypical of spontaneous speech, and will hardly appear in formal written texts (such as swear adjectives), so the lemmata selection will be affected. This aspect is related to a second important characteristic of the data: precisely the distribution of text types. Note the high influence of literature in the corpus, as opposed to other genres: 30.6% of the texts are literary, as opposed to for instance only 12.9% press.

Regarding the topic distribution, first note that the definition of topics is highly academic (Philosophy, Language and Literature, History and Geography. . .), with only a small proportion of non-academic concepts, such as press or leisure. Also note that, in terms of words, religion has a weight similar to that of press: 10.2% vs. 12.3% within non-literary texts, that is, religion occupies 7.1% of the overall corpus. This characteristic again affects lexical choice, so that adjectives related to the topics covered have a much higher frequency in the CTILC fragment than in other kind of corpora.

As an example, compare the rank of three religion-related adjectives in the CTILC fragment and in CUCWeb, a Catalan corpus built from the Web (Boleda et al., 2006)[3], as shown in Table 2.3.

adjective	translation	CTILC	CUCWeb
religiós	religious	99	251
catòlic	Catholic	218	358
eclesiàstic	ecclesiastical	603	998

Table 2.3: *Rank of three religion-related adjectives in two corpora.*

[2] This fact and the two examples that follow were brought to my attention by Gregory Guy, personal communication, October 2003.

[3] Interface available at `http://www.catedratelefonica.upf.es/cucweb`.

“Religious” adjectives are higher-ranked, and thus comparatively much more frequent, in CTILC than in CUCWeb. Of course, Web corpora are also biased towards certain kinds of texts (presumably, adjectives having to do with Computer Science are much more frequent in a Web corpus than in a traditional corpus). There is no such thing as a balanced corpus, because there are no universal balance criteria. However, given that all corpora are biased, a description of their content is essential to be able to place the results within a given context. In our case, we have to take into account that the kind of language we have access to through the corpus is formal written text belonging to literature and certain academic fields.

2.1.1.2 Linguistic processing

The CTILC corpus has been manually corrected, so that it is a resource with high-quality. However, it only provides lemma and morphological information. We used CatCG, a suite of Catalan processing tools developed at the GLiCom group (Alsina et al., 2002), to add a syntactic level of annotation to the corpus. CatCG comprises a tokenizer, a computational dictionary, a morphological tagger and a shallow parser.

We only needed the last tool, because all the previous steps were already covered in the resource. However, we had to adapt many formatting aspects of the corpus to be able to add syntactic infromation. The steps we followed were:

- adapt the segmentation of the corpus data to meet the format of CatCG
- enhance the computational dictionary with missing lemmata (this step was not strictly necessary, but was done for completeness)
- translate the morphological tagging used in the CTILC to the tagging used in CatCG
- parse with the shallow parser

We thus obtained a disambiguated, manually corrected morphological tagging in the GLiCom format, so as to be able to parse the corpus. More details on the kind of information encoded in the corpus and the CatCG tools can be found in Section 2.2.1.

2.1.2 Adjective database

Roser Sanromà developed a database of Catalan adjectives (Sanromà, 2003). She selected a particular set of lemmata from a corpus and manually encoded morphological information with respect to derivational type (whether an adjective is denominal, deverbal or not derived) and the specific suffix in case the adjective is derived. She included 2540 adjectives in the database.

She initially extracted all adjectives that appeared at least 25 times in a smaller fragment of CTILC (8 million words) than the one we have considered. The list obtained was pruned according to linguistic criteria. She manually coded the derivational type of the adjective (denominal, not derived, participial, or deverbal), as well as the suffix of the derived adjectives. For more details on the selection criteria and the information included in the database, see (Sanromà, 2003).

In initial experiments (see Chapter 5) with a 10 frequency threshold for adjectives, it was clear that the data were too sparse for results to be useful. We decided to establish a higher threshold, requiring at least 50 occurences, and selected from Roser Sanromà's database all those adjectives that occured more than 50 times in our fragment of the CTILC. Because the fragment is almost double than the one Roser Sanromà used, we only lost 10% of the data when doubling the threshold. The final database we used consisted of 2291 lemmata.

Note that this list includes some gerunds and participles with a predominantly modifying function. In GLiCom's computational dictionary, participles are only encoded as verbs, not as adjectives. This decision is based on two arguments: first, the high complexity (even for humans) in disambiguating between the two categories (Marcus et al., 1993; Jurafsky and Martin, 2000; Brants, 2000). Second, participles share the functions of adjectives and verbs.

Therefore, the verb/adjective ambiguity for participles is partially solved at the syntactic level: verbal participles carry out the same functions as any verb form, while adjectival participles the same functions as any adjective. The only distinction that is not solved is the most difficult one, namely, distinguishing between a verbal participle and an adjectival passive when they work as noun complements. See Section 3.1.3 for further discussion on delimitation problems.

For our experiments, we considered a participle as an adjective if it had any of the syntactic functions of an adjective, namely, pre- or post-nominal modifier, predicate of a copular sentence, or predicative complement.

2.1.3 Adjectives in the working corpus

This Section describes the distribution in the corpus (Section 2.1.1) of the lemmata included in the adjective database (Section 2.1.2). The general frequency distribution of adjectives in the corpus is, as expected, Zipfean, as shown in Figure 2.1. [4]

This distribution, where most items are very unfrequent but some items are extremely frequent, is typical of many linguistic phenomena. Table 2.4 presents this effect in a more concrete way, showing that two thirds of the adjectives in the database have less than 250 occurences (17.2 per million), while less than 8% have more than 1,000 occurences. Only 15 adjectives have more than 5,000 occurences. The three most frequent adjectives in the database are *bo* ('good'; 12,650 occurences), *nou* ('new'; 13,500 occurences), and *gran* ('big, great'; 19,551 occurences).

Frequency	lemmata	%lemmata	cumulative	%lemmata
50-99	739	32.2	739	32.2
100-249	800	34.8	1539	67.0
250-499	332	14.4	1871	81.4
500-999	238	10.4	2109	91.8
1000-5000	174	7.6	2283	99.3
>5000	15	0.7	2291	100

Table 2.4: *Frequency distribution in numbers.*

These frequencies are not evenly distributed. Morphology plays a major role in explaining

[4] The two graphics in the lower part of the figure are boxplots. The rectangles have three horizontal lines, representing the first quartile, the median, and the third quartile, respectively. The dotted line at each side is at most 1.5 times the length of the rectangle, and values that are outside this range are represented as points (Verzani, 2005). We will encounter this kind of representation many times in the thesis, as it is a complete representation of the distribution of feature values for a variable.

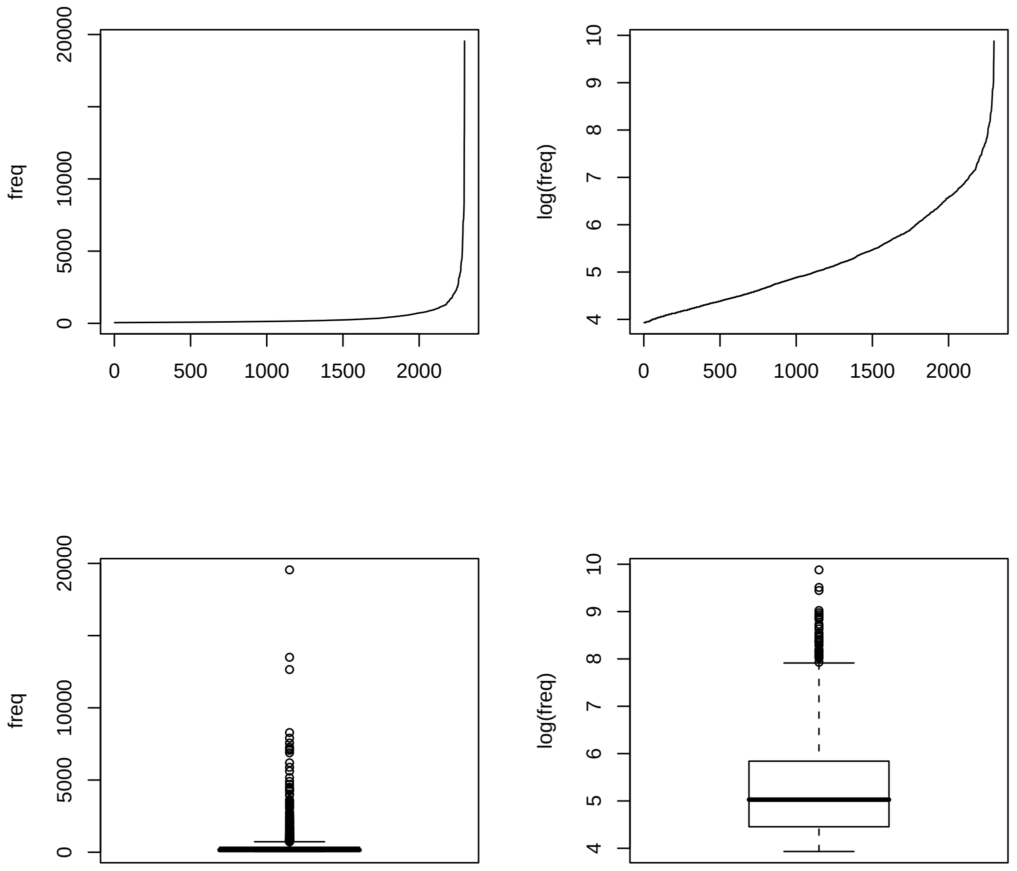


Figure 2.1: *Frequency distribution of occurences.*

frequency distribution of adjectives, as will be further explored in this thesis. Table 2.5 summarises the number of lemmata and total occurences represented by each of the morphological types coded in the database.

Table 2.5 shows that there are many more denominal lemmata (types) than any of the other types (37.5%). The ratio for the number of occurences (tokens), however, is similar to the ratio for the number of lemmata, if a bit lower (33.9%), which shows that on average denominal adjectives tend to have a lower number of occurences than the rest of adjectives. The tendency is the reverse for not derived adjectives: they cover 37.5% of the occurences, while only having 22.6% of the lemmata.

Deverbal and participial adjectives have an even greater difference between their type and token rations than denominal adjectives: 22.5% lemmata and 16.1% occurences for participial adjectives, and 17.4% lemmata (the smallest class) and 12.5% occurences for other deverbal adjectives. Thus, on average deverbal and participial adjectives are the least frequent classes. The means depicted in the last column of the table confirm this analysis.

The data in Table 2.5 indicate that derived adjectives in Catalan are more numerous (number of

Morph. type	type	%type	token	%token	mean
denominal	860	37.5	310,524	33.9	361.1
not derived	515	22.5	342,365	37.4	664.8
participial	517	22.6	148,017	16.2	286.3
deverbal	399	17.4	114,255	12.5	286.4
total/general	2291	100	920,821	100	152.5

Table 2.5: *Distribution of lemmata and occurences according to morphological type.*

lemmata) but less frequent (number of occurences). Conversely, not derived adjectives are less numerous but more frequent.

If one accepts that more frequent elements of a category are more prototypical, not derived adjectives are the core of this part of speech. This explains the greater attention that not derived adjectives (particularly some prototypical classes, such as size, colour, and temperature adjectives) have received in the literature, as opposed to derived adjectives (see Chapter 3).

However, the large number of derived lemmata in Catalan and other Romance languages warrants more attention to the semantics of these adjectives. Note that Dixon (2004) claims that across languages, “Typically, a higher proportion of adjectives than of nouns and verbs will be derived forms”.

The tendency shown in Table 2.5 would probably be more extreme if lesser frequent adjectives were taken into account, as the 50-occurence threshold represents 3,5 occurences per million word, which is not very low.

Details on the distribution of particular suffixes within each derivational class will be given in Chapter 4 (Section 4.2.1).

2.2 Tools

2.2.1 Shallow parser

As explained in Section 2.1.1, the corpus was parsed with the shallow parser included in CatCG. It is a manually developed grammar written in the Constraint Grammar formalism (Karlsson et al., 1995), and compiled with Connexor’s commercial tools[5] in runtime.

The parser is a functional shallow parser, that is, it assigns function tags to words without actually building a tree structure. For instance, a noun may be tagged as Subject, but no indication of which verb it is the subject of is provided, even if there is more than one finite verb in the sentence (as is case in e.g. embedded clauses).

For some functions, partial structural information is provided. One of the prototypical examples is the nominal modifier function for adjectives. CatCG will indicate an adjective is a pre-nominal modifier or a post-nominal modifier, but not the actual head. For instance, in both examples (2.1a) and (2.1b) the adjective *blanc* would be tagged as postnominal modifier, although in example (2.1a) it modifies *torre* and in (2.1b) it modifies *marbre*.

In this case, morphology is a clear indicator of the head: *torre* is a feminine noun, so that

[5] `http://www.connexor.com/`

blanc gets the feminine agreement suffix *-a* and is thus realised as *blanca* in (2.1a). Conversely, *marbre* is a masculine noun, so that *blanc* appears in the masculine form in (2.1a). However, if both nouns were masculine, it would not be possible to disambiguate the head, and indeed CatCG does not provide any clue with respect to the head other than the direction in which it should be looked for (left or right of the adjective).

(2.1) a. una torre de marbre **blanca**
a tower of marble white

‘a white tower of marble’

b. una torre de marbre **blanc**
a tower of marble white

‘a tower of white marble

Figure 2.2 shows an example of the kind of information available in the corpus that was used to model the data for the experiments explained in Chapters 5 and 6. [6]

Word	**Trans.**	**Lemma**	**Morph. info**	**Function**
El	the	el	masc. sg. article	pre-head det.
llenguatge	language	llenguatge	masc. sg. common noun	subject
d’	of	de	preposition	post-head mod.
aquests	these	aquest	masc. pl. demonstr. det.	pre-head det.
versos	verses	vers	masc. pl. common noun	post-preposition mod.
és	is	ser	3rd. p. sg. pres. ind. verb	main verb
poc	little	poc	adverb	pre-adjective mod.
respectuós	respectful	respectuós	masc. sg. adjective	predicate
.	.	.	full stop	punctuation

Figure 2.2: *Extract of the CTILC corpus.*

The annotation provided by GLiCom’s shallow parser is, thus, a kind of subspecified dependency grammar. Function tags either indicate a traditional syntactic function, such as subject or main verb, or the part of speech of the head of the phrase a word is in. For instance, *poc* is tagged as an adjective modifier, which indicates that it structurally belongs to the phrase headed by an adjective. Because the adjective *respectuós* functions as a predicate, and *poc* depends on the adjective, *poc* belongs to the predicate. However, recall that only the direction of the head is indicated (*poc* is a pre-adjective modifier, thus its head is to be found to its right), so that if there were another adjective with another function it would not be possible to identify the head of the adjective modifier without further processing.

[6] Legend:

masc.	masculine
sg.	singular
pl.	plural
demonstr.	demonstrative
det.	determiner
mod.	modifier
p.	person
pres.	present
ind.	indicative

The kind of annotation provided by CatCG of course affects the kind of information that can be extracted from the corpus. For instance, we can recover information that approximates nonrestrictivity (see Section 3.6.1.1), because we know whether an adjective is a pre-nominal or a post-nominal modifier. However, we can only identify its head with heuristics, so that we cannot estimate the distance to the head (information about adjacency; see Section 3.6.1.3) in a reliable manner.

CatCG does not always provide fully disambiguated output: there usually remains some degree of syntactic ambiguity in cases where the rules that apply cannot uniquely identify an appropriate function tag. This property increases the noise caused by automatic parsing. The syntactic properties of adjectives will be further explored in Sections 3.1.2 and 3.6.1.

2.2.2 CLUTO

The clustering tool used for the experiments reported in Chapter 5 is CLUTO (Karypis, 2002), a freely available[7] software package. It is highly parametrisable with respect to clustering method, similarity metric, clustering criterion, and other parameters of the clustering procedure. More details about the options used will be offered in Section 5.1.3, which provides an overview of the clustering technique.

CLUTO also provides facilities for exploration and understanding of the results. On the one hand, information of each cluster with respect to size, properties of the cluster (similarity of objects among them and with respect to other clusters), and predominant features. On the other, several graphics that allow the user to visualise the structure of the clustering and the distribution of feature values. The handbook for CLUTO (Karypis, 2002) contains quite detailed documentation of all the aspects mentioned. Recently, a freely available graphical version of CLUTO, gCLUTO, has been developed.

2.2.3 Weka

The tool used for the experiments reported in Chapter 6 is Weka (Witten and Frank, 2005), a freely available[8] Java software package with several graphical interfaces. It includes most of the algorithms used nowadays in the Machine Learning community for classification, from Bayesian classifiers to ensemble classifiers, as well as several clustering algorithms. It also offers feature selection and association algorithms, and some visualisation facilities for input and output data.

Weka can be used as a stand-alone tool, to perform individual experiments (Explorer interface) or massive experiments (Experimenter interface), or its functions can be called from a console or embedded in the user's code. It also facilitates evaluation, offering a range of evaluation procedures (cross-validation, train/test, holdout method) and evaluation metrics (accuracy, information theory measures, kappa, per-class recall and precision, etc.). Its Experimenter interface provides a significance test for comparing the performance of two algorithms.

[7] At `http://glaros.dtc.umn.edu/gkhome/views/cluto`.

[8] At `http://www.cs.waikato.ac.nz/ml/weka/`.

2.2.4 Other tools

The architecture of the experiments (Chapters 5 and 6), from data extraction to file management, was developed using the Perl programming language[9]. Its CGI module was used in the implementation of the Web experiment. The data obtained during the Web Gold Standard experiment (Chapter 4) was stored and processed in several ways within a MS Access database. For most statistical analysis purposes, implementation of agreement measures and graphic production, the open source R language and environment was used (R Development Core Team, 2006).[10].

[9] See e.g. `http://perldoc.perl.org/`

[10] Homepage: `http://www.r-project.org/`

Chapter 3
Adjective semantics and Lexical Acquisition

> From the beginning, I concentrated on the understanding of *words*: not words such as ‘all’, ‘and’, and ‘necessarily’, but rather words such as ‘yellow’, ‘book’, and ‘kick’.
>
> Marconi (1997, p. 1)

This chapter is devoted to theoretical considerations about adjective classification. The literature on adjectives is scarcer than literature on nouns, and much scarcer than literature on verbs, both in theoretical and computational linguistics. In addition, there does not exist a single, well-established classification proposal for adjectives. We reviewed the literature and performed empirical research to define a classification that could be adequate for our purposes. Such a classification is subject to a series of constraints, that we review before entering the discussion of the literature.

General constraints The classification should be **broad**, that is, its parameters should cover most adjectives. This excludes for instance an organisation in terms of antonymy such as that in WordNet, for we have seen in Section 3.5.1, that this property is only applicable to a small subset of adjectives. It should also be **consistent**, which means that its parameters should be established at more or less the same level of description for all adjectives. The classification proposed within the SIMPLE project (Section 3.5.2), which mixes formal semantics and denotational categories, is prone to problems where the two parameters do not have a natural fit. Another inconsistent classification is the one established in some descriptive grammars (Section 3.2), which mixes morphological, syntactic and semantic criteria.

Finally, the classification should be **balanced**, in the sense that the classes should not be enormously different in size. From the perspective of Natural Language Processing, the particularities of a handful of adjectives can be manually established, so that there is no need for automatic acquisition. From a theoretical perspective, if one of the parameters used in the classification distinguishes a very small subset of adjectives from a large heterogeneous body of other kinds of adjectives, its usefulness to draw central distinctions among their semantics can be questioned.

Linguistic constraints The first linguistic constraint requires the classification to be of a semantic nature, that is, classes should define broad, core semantic characteristics. However, the characteristics should have correlates in other levels of description, such as morphology and syntax. That constraint provides an empirical handle for semantic description (contrary to purely denotational approaches such as Dixon (1982)), and a suitable testbed for competing theories. It also allows linguistic theory and NLP systems to exploit the **morphology-semantics** and **syntax-semantics** interfaces for adjective classification, establishing regularities at morphology and syntax that provide insights into semantics.

Related work in Lexical Acquisition for English and German verbs (Merlo and Stevenson, 2001; McCarthy, 2001; Korhonen, 2002a; Schulte im Walde, 2006) has exploited the syntax-semantics interface. The idea is to test to what extent syntactic differences correspond to semantic differences, under the assumption that semantically similar predicates will tend to behave similarly in relevant syntactic aspects. This hypothesis is commonly attributed to Harris (1968), and has led to fruitful insights in research on lexical semantics (see, e.g., Levin (1993) on verb classes).

The morphology-semantics interface is not as "popular" in linguistic theory and NLP as the syntax-semantics interface. However, the relationship between morphological processes and regular meaning effects has been repeatedly noted since the beginning of studies on language. Some morphological processes, such as derivation, have a clear and relatively stable semantic effect.

These interfaces can be exploited in two directions: to induce semantic properties from morphological and syntactic properties, and to exploit semantic properties to predict linguistic behaviour. The general goal of predicting properties at one level from properties at another level is common to linguistic theory.

NLP constraints The results of the research presented here are intended to be used for NLP purposes, which imposes further constraints on the task. Note that one of the main constraints within NLP is again coverage, because the classification should apply to all lemmata contained in a particular computational dictionary. This aspect has been mentioned above as a general constraint. Also, in NLP, higher priority is assigned to higher frequency phenomena, due to the difference in goals with theoretical linguistics (Nirenburg and Raskin, 2004, p. 111).

The concrete setting of the NLP resources available for a particular language and the system used also have to be taken into account. CatCG, the NLP system for Catalan developed at the GLiCom group (see Section 2.2.1), is symbolic in nature. In addition, symbolic knowledge is interpretable and can be manipulated by trained linguists, so that automatically acquired information can be checked and enhanced. We therefore aim at symbolic classes rather than, e.g., probabilities across patterns. Probabilistic information can be added in subsequent modules or versions of the system.

A second main constraint imposed by the NLP setting is the use of automatic means to induce classes (as opposed to manual classification) and, relatedly, a limitation on the amount and kind of information that can be used to perform automatic acquisition. A particular corpus (Section 2.1.1) will be used, annotated with morphological and shallow syntactic information. An additional source of information is a database of adjectives (Section 2.1.2) with information on derivational morphology. This constraint relates to the requirement about exploiting interfaces between levels of linguistic description mentioned above: the requirement is theoretically motivated, but also methodologically convenient.

These constraints were born in mind in assessing the proposals from several traditions. In what follows, we first summarise the characteristics of adjectives in Catalan as a part of speech (Section 3.1) and then review the treatment of their semantics in different fields: descriptive grammar (Section 3.2), formal semantics (Section 3.3), ontological semantics (Section 3.4), and Natural Language Processing resources (Section 3.5). Section 3.6 discusses the kind of classification that is aimed at in this PhD, and how different aspects of the reviewed research fit into the desired classification. Section 3.7 provides a review of previous work on Lexical Acquisition for adjectives. The chapter ends with a general summary (Section 3.8).

3.1 Adjectives as a part of speech

Adjectives can be defined as a morphosyntactic class which in Catalan and other languages present a set of relatively homogeneous morphological and syntactic characteristics (see Picallo (2002) for Catalan, Demonte (1999) for Spanish, and Huddleston and Pullum (2001) for English; for a typological perspective, see Dixon and Aikhenvald (2004)). This section reviews their main morphological and syntactic characteristics, as well as problems in delimiting the class of adjectives from that of nouns and verbs, focusing on Catalan.

3.1.1 Morphology

The main morphological characteristics of adjectives in Catalan are summarised in what follows (see Badia i Margarit (1995) and Picallo (2002)).

Inflection: adjectives inflect for number and most of them for gender: *econòmic, econòmica, econòmics, econòmiques* (‘economic(al)’) vs. *alegre, alegres* (‘joyful’). They do not inflect for tense or mood, contrary to verbs.

Agreement: their inflectional properties in a particular occurrence depend on the head noun or determiner they modify, that is, they do not have inherent gender or number features, but agree with a nominal head with respect to gender and number.

Derivation: there are many deverbal and denominal adjectives: *abusiu* (‘abusive’; from *abusar*, ‘abuse’), *econòmic* (‘economic(al)’; from *economia*, ‘economy’). Many adjectives admit derivation through the degree suffix *-íssim* (*netíssim*, ‘very clean’), though not all (**agricolíssim*, ‘very agricultural’). Those which admit *-íssim* generally also admit diminutive suffixes such as *-et* or *-ó*: *netet, petitó* (‘clean$_{diminutive}$’, ‘little$_{diminutive}$’).

Composition: they can form so-called *copulative composts* (Picallo, 2002, 1645), that is, compounds obtained by juxtaposition of two adjectives: *clarobscur*, ‘light-dark’, *agredolç* ‘sweet-sour’, and intervene in a productive Noun + Adj composition process, where a notion of inalienable possession is necessary: *culgròs* ‘big-bottom’, but not **samarreta-gròs*, ‘big-T-shirt’.

Openness: they are an open (productive) class of words, mainly through the derivation and (to a lesser extent) composition processes mentioned in the previous paragraphs.

3.1.2 Syntax

Adjectives have two main syntactic functions: they act as predicates or noun modifiers. Even though more detailed syntactic environments can be identified (Yallop et al. (2005) identify over 30; see Section 3.1.2), they are specialisations of these two functions.

The default function of the adjective in Catalan is that of modifying a noun; the default position is the postnominal one, as in most Romance languages Picallo (2002). Examples are *taula gran*, 'big table', *arquitecte tècnic*, 'technical architect' and *element constitutiu*, 'constitutive element'. However, some adjectives can appear prenominally, mainly when used non-restrictively (so-called "epithets"; see Section 3.6.1.1). *Gran taula* 'big table' is possible, but not **tècnic arquitecte, *constitutiu element.*

Adjectives can also function as predicates, be it in a copular sentence (example (3.1a)) or in other predicative contexts, such as adjunct predicates (example (3.1b)).

(3.1) a. Aquest cotxe és molt maco.
This car is very nice.

'This car is very nice.'

b. la vaig veure borratxa
her PAUX-1ps see drunk

'I saw her drunk'

Table 3.1 lists the main syntactic functions for adjectives together with their distribution in the corpus. Note that post-nominal modification covers an overwhelming majority of the cases (62.7%), and that predicative functions are much less frequent than nominal modification (21.7% opposed to 78.3% of the cases). The latter fact is to be expected from the formal register and written form of most texts in the corpus used, as mentioned in Section 2.1.1.1.

Function	Cases	Subfunction	Cases	%Total
Modifier	927,203	Post-nominal modifier	742,375	62.7%
	(78.3%)	Pre-nominal modifier	184,828	15.6%
Predicate	256,467	Copular sentence	129,531	10.9%
	(21.7%)	Other environments	126,936	10.7%

Table 3.1: *Distribution of adjective syntactic functions in the CTILC.*

Complements to adjectives are expressed as PPs or clauses, and the preposition is dropped when the complement is expressed as a clause, as exemplified in (3.2).

(3.2) a. **Orgullós** de la seva filla
Proud of the his daughter

'Proud of his daughter'

b. **Orgullós** que la seva filla treballi molt
Proud that the his daughter works much

'Proud that his daughter works hard'

Adjectives rarely have more than one complement, but there are some cases (p. 1660, Bonet and Solà (1986)], taken from p. (2002)):

(3.3) a. **Dependent** de la seva dona en moltes coses
Dependent of the his wife in many things

'Dependent on his wife in many aspects'

b. **Superior** a en Joan en matemàtiques
Superior to the Joan in maths

'Superior to Joan in maths'

Adjectives are modified mainly by adverbs: degree adverbs (*tan rica* 'so rich', *massa net*, 'too clean') or other kinds of adverbs (*predominantment agrícola*, 'mainly agricultural').

3.1.3 Delimitation problems

Adjectives as a part of speech can not always easily be distinguished from nouns and verbs. As for the adjective-noun distinction, in some cases human-referring adjectives "act" as nouns, as can be seen in examples in (3.4):

(3.4) a. The rich and the poor

b. El cec va marxar.
The blind PAUX-3ps go

'The blind (man) went away.'

Whether in these cases *rich* or *cec* are morphologically nouns (have undergone conversion) or adjectives acting as nouns is subject to debate. In Catalan and other Romance languages, the difficulty is greater than in English, because of the general possibility of "headless noun phrases" (again, whether there is a phonologically null head in these constructions is a matter of debate):

(3.5) a. Havien buidat les dues ampolles, **la de xerès** i **la de conyac**
Had emptied the two bottles, the of sherry and the of cognac

'They had emptied both bottles, the sherry bottle and the cognac bottle.'

b. La Bruna vol el cotxe verd, però jo prefereixo **el blau**.
Bruna wants the shirt green, but I prefer the blue.

'Bruna wants the green shirt, but I prefer the blue one.'

c. **El blau** és el meu color preferit.
The blue is the my colour favourite.

'Blue is my favourite colour.'

In example (3.5a), the two headless NPs contain a determiner and a prepositional phrase headed by *de* ('of'). In example (3.5b), the headless NP contains a determiner and an adjective. Note that it is superficially identical to the first NP in example (3.5c), in which *blau* is arguably a true noun. This case is quite uncontroversial, but there are cases that are not so clear-cut, like the following (all adapted from CTILC):

(3.6) a. De Gaulle no podia veure homes com Picasso o Jean-Paul Sartre, perquè eren comunistes
De Gaulle not could see men like Picasso or Jean-Paul Sartre, because were **communists**

Category	#Lemmata	%
Adjective or noun	5532	32%
Total adjectives	17281	100%

Table 3.2: *Adjective-noun ambiguity in GLiCom's computational dictionary.*

'De Gaulle could not stand men like Picasso or Jean-Paul Sartre because they were communists'

b. hauria d'esser una relació d'**igual** a **igual**
should of-be a relationship of-equal to equal

'It should be an equal-to-equal relationship'

c. era un **entusiasta** universitari a la seva manera
was an enthusiast of-university to the his way

'He was a university enthusiast in this way'

d. he parlat de la qüestió amb **naturals del país**
have talked of the question with native-people of-the country

'I have talked about this question with people from the country.'

e. Els cabarets [són] ocupats per milicians, no pas per **joves burgesos**
The cabarets [are] occupied by soldiers, not NEG by young bourgeois

'Cabarets are filled with soldiers, not with young bourgeois men.'

For instance, it is not easy to determine whether *naturals* is an adjective or a noun in example (3.6d), or whether in (3.6e) the head is *jove* or *burgesos*. Carvalho and Ranchhod (2003) discuss a very similar example in Portuguese: *uma joven cigana* could be a young woman that is gipsy or a gipsy woman that is young. Note that in all these cases the NPs refer to humans. Inflective properties do not help in these cases, because although nouns exhibit fixed gender in Catalan, for human referring nouns there are usually both feminine and masculine versions of the nouns (e.g., *nen* for 'boy' and *nena* for 'girl').

As can be seen in Table 3.2, 32% of adjective lemmata in GLiCom's computational dictionary have an additional nominal reading. Of course, in many cases the distinctions are not controversial for humans (are rather of the (3.5b-3.5c) type). However, this represents a huge problem for NLP purposes. The same problem arises in other Romance languages, for instance Portuguese (Carvalho and Ranchhod, 2003). Also in Germanic languages the delimitation of the nominal and adjectival category causes problems in automatic and manual annotation. Brants (2000) shows that nouns and adjectives are involved in 42.2% of the differences between two human annotations of a German newspaper corpus, and the third pair with highest-ranked confusion rate is the noun-(attributive) adjective pair.

The other edge is represented by the adjective-verb distinction, for verb forms that in Romance languages share inflection with adjectives: past and present participle. A gradation from more verbal to more adjectival constructions in which past participles take part is given in examples (3.7a-3.7e).[1]

[1] We will concentrate on past participles, rather than present participles, as the former are most productively used as adjectives. There are 513 past participles as opposed to 140 present participles in the adjective database used for this PhD (see Section 2.1.2).

(3.7) a. Algú ha abatut el soldat
Someone has shot-down the soldier

‘Someone has shot the soldier down’

b. El soldat ha estat abatut per l’enemic
The soldier has been shot-down by the-enemy

‘The soldier has been shot down by the enemy’

c. El soldat ha estat abatut
The soldier has been shot-down

‘The soldier has been shot down’

d. Demà enterren el soldat abatut per l’enemic.
Tomorrow bury the soldier shot-down by the-enemy

‘The soldier that was shot down by the enemy will be buried tomorrow’

e. Demà enterren el soldat abatut.
Tomorrow bury the soldier shot-down

‘The shot-down soldier will be buried tomorrow’

Example (3.7a) is a fully verbal construction. Example (3.7b) is a clear passive (note the agentive *by*-phrase). Example (3.7c) could be a passive or a copular sentence, as in *El soldat ha estat molt trist últimament* ‘The soldier has lately been very sad’. Examples (3.7d) and (3.7e) represent exactly the same dichotomy in the nominal, as opposed to verbal, domain: in Catalan, a past participle occuring within a noun phrase can occur with an agentive *by*-phrase, as in 3.7d).

In many cases, a past participle used as an adjective undergoes a shift in meaning, thus arguing for an adjectival entry in addition to the verbal entry. In Catalan, *abatut* also means ‘depressed, blue’. Therefore, sentences (3.7c) and (3.7e) are ambiguous out of context. In sentences (3.8a) and (3.8b), parallel to (3.7c) and (3.7e), *abatut* is unambiguously used as an adjective.

(3.8) a. El soldat ha estat molt abatut últimament
The soldier has been very depressed lately

‘The soldier has lately been very depressed’

b. Aquell soldat tan abatut em va fer pena.
That soldier so depressed me PAUX-3ps do sorrow.

‘I felt pity for the very depressed soldier.’

(3.9) a. #El soldat ha estat molt abatut per l’enemic
The soldier has been very shot-down by the-enemy

‘The soldier has been shot down by the enemy’

b. #El soldat ha estat abatut per la malaltia de la seva mare
The soldier has been very depressed by the illness of the his mother

‘The soldier has been depressed by the illness of his mother’

Category	#Lemmata	%
Adjective or verb	3934	16.7%
Total adjectives	23591	100%

Table 3.3: *Adjectival and verbal participle forms in the CTILC corpus.*

c. El soldat ha estat abatut a causa de la malaltia de la seva mare
The soldier has been depressed to cause of the illness of the his mother

'The soldier has been depressed because of the illness of his mother'

Note that in this use, *abatut* is gradable, which it is not when acting as a true participle (3.9a). Also, it does not admit a *by*-agentive phrase (3.9b), although it does admit other causal adjuncts (3.9c). The phenomenon of so-called 'adjectival passives' such as in (3.7e) has been well studied in English (Bresnan, 1982; Levin and Rappaport, 1986; Bresnan, 1995). Prefixation with *un*, prenominal modification, modification with *too* or *too much*, complementation with a direct object, and heading of concessional relative phrases beginning with *however* are behaviours that distinguish adjectival from passive uses of the past participle according to Bresnan (1995).

However, even with these tests, the fact remains that the semantic distinctions are very subtle, and their syntactic distribution very similar. This characteristics makes it difficult to treat the distincion with automatic means, and also causes it to be controversial for humans as described in the literature (Marcus et al., 1993; Jurafsky and Martin, 2000; Brants, 2000). In fact, in the corpora such as British National Corpus and the Penn TreeBank, underspecified tags are used to tag participial forms so as to bypass this difficult matter.

In the GLiCom computational dictionary, an even more radical approach is taken: participles are uniformly tagged as verbal forms. The adjective/verb ambiguity is implicitly solved at the syntactic level, in deciding whether they act as nominal modifiers or main verbs, for instance. The manual tagging of the CTILC corpus provided by the Institut d'Estudis Catalans does distinguish between adjective and participle functions.

Based on the manual tagging, the scope of this ambiguity can be determined. Table 3.3 shows that 16.7% of the adjective lemmata that occur in the CTILC are participles that can be considered to be lexically ambiguous between an adjectival and verbal use.

In the research undergone here, these criteria were approximated, because CatCG does not distinguish between adjectival and verbal readings of participles. Participles with an adjectival syntactic function (noun modifier, predicate in a copular sentence or predicate in other constructions) were considered to be instances of adjectives. Participles bearing verbal syntactic functions were not taken into account for the experiments.

3.2 Adjectives in descriptive grammar

In this section, we follow the discussion in Picallo (2002). It is representative of the kind of treatment adjectives undergo in descriptive grammars, and facilitates the introduction of relevant notions for the characterisation of adjectives.

The definition of adjectives as a part of speech in Picallo (2002, p. 1643) is based on semantic and pragmatic properties:

> The adjective is the part of speech which is used to attribute properties to an entity or to classify , in different ways, concrete or abstracts objects, events, processes, states, or actions which nouns denote. One of the main objectives of the use of the adjective is to contribute to delimit the class or kind of entity that the noun designates (Kamp 1975:153). It can also contribute to distinguish a particular entity among all the others in its same class or kind. [2]

She proposes a three-way classification:

- qualitative (*qualificatius*)
- relational (*de relació*)
- adverbial (*adverbials*)

Picallo states that the classification is based both in interpretative and syntactic features, thus implying that the syntax-semantics interface is at play in the classification. However, the relationship between semantic and syntactic characteristics remains quite implicit in the discussion except for some concrete properties. The relationship between syntax and semantics for adjectives will be discussed in Section 3.6.1. We now turn to discussing the classes proposed by Picallo in more detail.

3.2.1 Qualitative adjectives

Qualitative adjectives "name properties of the entities that nouns denote. With an adjective, one or more qualities of an object are expressed, which, in a particular situation or context, the speaker wants to distinguish among many other actual or possible possibilities." Picallo (2002, 1646) [3]

The definition of the class is again semantic ("properties of entities") and pragmatic ("the speaker wants to distinguish"). Some of the examples she provides are *benèvol* 'benevolent', *triangular* 'triangular', *resistent* 'resistant, strong', *blanc* 'white', *agradable* 'nice'.

Picallo further proposes a subclassification of qualitative adjectives on denotational grounds, based on Dixon (1982) (Picallo, 2002, 1648-1651): dimension, physical properties, speed and position, age, colour and shape, character, and evaluation. Again, she states that this division allows for a description of morphosyntactic and lexical properties. However, she only discusses adjective ordering. According to her, the following groups account for the ordering of adjectives in Catalan (ordering according to closeness to the head noun):

group I colour and shape adjectives (in that order)

[2] "L'adjectiu és la part de l'oració que s'empra per atribuir propietats a una entitat o per classificar de maneres diverses els objectes concrets o abstractes, els esdeveniments, els processos, els estats o les accions que denoten els noms. Un dels objectius principals de l'ús de l'adjectiu és contribuir a delimitar la classe o el tipus d'entitat que el nom designa (Kamp 1975:153). També pot contribuir a distingir una entitat particular d'entre totes les altres de la seva mateixa classe o tipus." My translation (GBT).

[3] "Els adjectius qualificatius anomenen propietats de les entitats que denoten els noms. Amb l'adjectiu s'expressa una o diverses qualitats d'un objecte que, en un context o situació determinats, el parlant vol distingir d'entre moltes altres possibilitats actuals o possibles." My translation (GBT).

group II polar adjectives (age, physical properties, dimension and speed and position)

group III character and evaluation adjectives

She claims that the hierarchy states a neutral order which can be altered, but does not provide any empirical evidence nor discusses the kind of mechanisms that alter order. Although order considerations will affect our classification (see Section 3.6.1.3), only inter-class, not intra-class ordering constraints will be taken into account.

This subclassification raises a further issue that will arise again in the discussion, namely, coverage. Dixon's classification covers the most prototypical adjectives, but does not account for the rest of the lemmata. For instance, adjectives such as *esquerre* ('left'), *sublim* ('sublime'), *total* ('total') can not be placed in this classification. A richer classification is needed, as will be argued in Section 3.4.

The main feature that characterises qualitative adjectives according to Picallo is gradability. However, not all qualitative adjectives are gradable (an exception Picallo mentions is *triangular*). This raises the issue of whether it is a property of the class or just of some items.

Relatedly, Picallo states that qualitative adjectives typically have antonyms and are organised around a scale between two poles, or *polar* scale. A typical polar scale is the size scale, formed among others by *petit-mitjà-gran* ('small-medium-large'). Picallo mentions the *relativity* of the domain to which the adjective applies: a long box is not of the same size as a long street. This aspect was discussed by Kamp (1975) as the *vagueness* of the meaning of the adjective, who points out that not all referents can be placed in a particular position on the scale. Otherwise, sentences such as *This man is neither tall nor short* would be contradictory. Scalar adjectives are typically gradable (*very small/cold/short*), but again, not all scalar adjectives are straightforwardly gradable.

Colour or shape adjectives are the most cited non-gradable scalar adjectives. These are called *complementary* scales, because they do not consist of two poles, but a set of related, unordered items. Picallo states that exemples like those in (3.10), which are abundant and natural, are not counterexamples, although she does not provide further arguments.

(3.10) a. La Maria té la cara més **rodona** que la Lluïsa
The Maria has the face more round than the Lluïsa

'Maria has a rounder face than Lluïsa'

b. els ous eren ben blancs.
the eggs were well white.

'The eggs were completely white.'

Scalar properties of complementary scales are different to polar scales. For typical scalar adjectives, augmented degrees of a position in a scale lead to the following position (*very hot → burning*). Because colour and shape adjectives are not ordered, it seems that what is being graded with the degree adverbs is the quality itself (roundness in example (3.10a, whiteness in example (3.10b), with no clear implications for the other elements in the scale. Note, however, that it is also possible to compare elements of the same scale, as in example (3.11)

(3.11) La Maria té la cara més rodona que quadrada
Maria has the face more round than square

'Maria's face is more round than square'

Data such as examples in (3.10) and (3.11) show that colour and shape adjectives are gradable, at least in Catalan.

The crucial question with respect to gradability and scalar organisation is to what extent these properties hold for the whole class of adjectives, or just for a couple of prototypical adjectives. The discussion about order (groups I, II, III) seems to indicate that not all adjectives are polar, only those in group II. As we will discuss in Section 3.5.1, most adjectives, even in the qualitative class are just not organised according to polar or antonymy relationships.

3.2.2 Relational adjectives

Relational adjectives have been tackled in Romance descriptive grammar at least since Bally (1944, §147, pp. 96-97), who called them *adjectifs de relation*.

Picallo characterises these adjectives by means of two properties, one syntactic, one semantic:

- they can only act as postnominal modifiers. This property will be examined in Section 3.3.2.1.
- they "do not express qualities that can be adscribed to an object in a maximal or minimal degree, but express entities. In [NPs such as *l'estructura molecular* 'the molecular structure'] the entity denoted by the nominal head is put in relationship with another entity denoted by the adjective" (Picallo, 2002, 1667).[4]

The view that relational adjectives denote objects, or are quite equivalent to nouns, is implicit in the term *pseudoadjectives* used by Postal (1969) to refer to this class, and has been given a derivational syntactic analysis in Levi (1978).

As will be explained in Section 3.3.2.1, we analyse relational adjectives as denoting **relationships** with a kind of "shadow argument" (Pustejovsky, 1995, pp. 62-67), the nominal embedded in the meaning of the adjective. This view is consistent with other studies on the semantics of relational adjectives (Fradin and Kerleroux, 2002; McNally and Boleda, 2004).

Due to this particular kind of denotation, according to Picallo, they are not gradable, so that they do not admit the superlative suffix *íssim*, nor degree modifiers. Also, they do not derive in *itat*, a suffix expressing property, which can be affixed to some qualitatives.

One of the issues about relational adjectives is their denominalness: is that a necessary and sufficient feature? Picallo says that most relational adjectives are denominal, with exceptions, but does not dwell further on the issue. There are denominal adjectives which are clearly intersective, such as *vergonyós* ('shy'; derived from *vergonya*, 'shyness'). Other adjectives are not (synchronically) denominal but truly relational: *bèlic* 'bellic, of-war', *botànic* 'botanical, of-plants', etc.

[4] "no expressen qualitats que es puguin adscriure a un objecte en grau màxim o mínim, sinó que expressen entitats. En [sintagmes nominals com *l'estructura molecular*] es posa en relació l'entitat que denota el nucli nominal amb una altra entitat que denota l'adjectiu" My translation (GBT).

Thus, denominalness is neither a necessary nor sufficient condition for a relational adjective. The crucial feature for this class is of a semantic nature, namely, the object-relationship denotation. The relationship between morphology and semantics will be taken up in Section 3.6.2. We defer the discussion of syntactic properties concerning relational adjectives to Section 3.6.1.

A final note about the semantics of relational adjectives: It has often been noted that relational adjectives can perform an argumental function, as in example (3.12a), which is semantically quite equivalent to "the Pope visits". In other cases, such as (3.12b), the relationship cannot be considered to be argumental.

(3.12) a. la visita papal
the visit Papal

'the Papal visit'

b. el cotxe papal
the car Papal

'the Papal car'

Picallo (see also Bosque and Picallo (1996)) treats this functional dychotomy in terms of **classes** of relational adjectives: thematic adjectives act as arguments (example (3.12a)), classificatory establish other kinds of relationships (example (3.12b). However, as has been shown in the examples in (3.12), some (if not all) relational adjectives can perform both functions, and the kind of relationship established depends on the head noun, not on the adjective.

3.2.3 Adverbial adjectives

The third class, adverbial adjectives, is very shortly discussed, and it is acknowledged that they are an "heterogeneous group of adjectives" (Picallo, 2002, 1683).[5] They are negatively defined, by neither expressing properties nor denoting entities. The only positive characteristic is that most of the adjectives in this class can enter an adverb deriving process with *-ment*.

Picallo discusses three subclasses:

Modal: They express notions related to modality. These are cases like *possible* 'possible', *necessari* 'necessary', *presumpte* 'alleged'. It is however problematic that *presumpte* is included in this group, as it does not relate to modality, but to speaker attitude (see Section 3.3.2.2). Modal adjectives modify propositions or events, rather than individuals.

Aspectual: They modify temporal semantic features: *els freqüents atacs romans a les tribus germàniques* 'the frequent Roman attacks to the Germanic tribes', *un visitant assidu* 'a frequent visitor'. These adjectives modify events (Larson (1998); see Section 3.3.2.1).

Circumstancial: "modifiers of the nominal entity which indicate, be it the situation of its referent (in the space or time), be it the manner in which an object is perceived or presented" (p. 1683)[6] For instance, *una ullada ràpida al manuscrit* 'a quick look at the manuscript', *l'avís previ* 'the previous warning', *la propera cantonada* 'the next corner' (Picallo, 2002, ex. 101, p. 1683).

[5] "grup heterogeni d'adjectius" My translation (GBT).

[6] "modificadors de l'entitat nominal que indiquen, bé la situació del seu referent (en l'espai o en el temps), bé la manera en què es percep o es presenta un objecte" My translation (GBT).

Adjectives in this class have the property that their arguments are not prototypical, individual-denoting arguments. Most of them typically modify propositions or event-denoting nouns, or have to do with indexical properties of the utterances (time, space).

3.2.4 Summary

The discussion of descriptive grammar has served to introduce relevant concepts regarding adjective classification. The morphology-semantics and syntax-semantics interfaces, implicitly used at several points in the discussion, are central to the work presented here. Denotational aspects (what kinds of meanings do adjectives have?) are also a central concern. Some of the distinctions will also be adopted, namely, the distinction between qualitative and relational adjectives. Adverbial adjectives will not be separately considered, but will be lumped together with qualitative adjectives, on the basis that they too denote attributes (even if they are not properties of individual, but of events or propositions) and that their syntactic behaviour is similar to qualitative adjectives (see Section 3.6.1).

3.3 Adjectives in formal semantics

3.3.1 Lexical semantics within formal semantics

Formal semantics as stemming from the Fregean philosophical tradition has concentrated on structural aspects of language meaning (such as logical connectors, tense, modality, polarity, or, above all, quantification), and has consciously neglected most of lexical semantics.[7] Because since Frege (1892) the main emphasis is placed on truth conditions, and the mechanisms for determining a truth value for a proposition expressed in a sentence from its parts (compositional semantics), “only issues that lend themselves to truth-conditional treatment are added to the inventory of formal semantic tasks” (Nirenburg and Raskin, 2004, p. 104).

Frege distinguished between two aspects of meaning: *sense* (Sinn) and *denotation* (*Bedeutung*). A famous example is the distinction between *the morning star* and *the evening star*. Both refer to the same celestial body (the planet Venus), but it can be that a person that knows that *Venus is the morning star* does not know that *Venus is the evening star*. Moreover, there is more to the meaning of *morning star* than the fact that it refers to (*denotes*) the planet Venus. It is its sense.

Richard Montague turned the philosophical work on the semantics of natural languages into a methodology for doing research in linguistics (his works are gathered in Thomason (1974)). This methodology is called model-theoretic semantics, because it is based on defining a model of the world and explicitly listing all the objects and relations holding among objects in that model.

The meaning of words is extensionally defined: the denotation of a particular noun, for instance, is the set of objects to which it can truthfully be applied. For instance, the denotation of *ball* is the set of objects in the model that are balls. The denotation of a particular transitive verb is the set of *pairs* of objects to which it can be applied. For instance, if John loves Mary and Mary loves Bill, the corresponding model specifies as denotation for the verb *loves* the pairs of individuals *{<john, mary>, <mary, bill>}*. If the set of pairs in the denotation of, e.g. *betrays* is the same as the set of pairs in the denotation of *loves*, there is no way to distinguish, within a

[7] For a review of the treatment of lexical semantics in philosophical semantics, see Marconi (1997, ch. 1).

particular model, the meaning of these two verbs.

From this very architecture it follows that formal semantics accounts for denotations or extensions, not for *senses* in Fregean terms. Moreover, senses are not even in its research program: "One of the radical aspects in Montague's work was the relegation of most of lexical semantics to a separate empirical domain" Partee (1996, p. 34).

One kind of device that was introduced to deal with this problem within philosophical semantics, and later imported into linguistic semantics, is the meaning postulate. A meaning postulate stipulates a relation between lexical items, or more precisely, between the denotations of the lexical items. For example, "all individuals that belong to the denotation of *bachelor* also belong to the denotation of *unmarried men*".

Meaning postulates cover the inferential aspect of meaning: assuming that a complete set of meaning postulates can be defined, we would be able to derive the validity of inferences such as that in example (3.13).

(3.13) a. These are cows

b. $\models$ These are animals

The main problem with meaning postulates, as Marconi (1997, p. 19) put it, is that "a meaningless (that is, uninterpreted) linguistic symbol cannot be made meaningful by being connected, in any way whatsoever, to more uninterpreted symbols. If you don't know Chinese, a Chinese definition for a Chinese word will not make you understand the word." Note that meaning postulates are relationships between predicates of a language, not between concepts.

Marconi also remarks that, from a cognitive point of view, it can be argued that we "know something about words such as 'book', 'table', and 'on' that is not captured by the meaning postulates for these words . . . What we know is, very briefly, how to apply such words in the real world." Inferential knowledge, although a part of lexical semantics, is not all there is to lexical semantics. In addition, it is not clear that meaning postulates are the best way to establish or represent inferential properties of words: we will see an alternative in Section 3.4.

Formal semantics is generally not adequate to treat many aspects of lexical semantics. In this framework, the meaning of words is typically an atom[8]. How does this general inadequacy reflect in the case of adjectives? This is the topic of the next Section.

3.3.2 Adjective classes from entailment patterns

Partee (1996, p. 34) precisely explains the treatment of adjectives as an "example of the kind of lexical semantics" that is done in formal semantics.

The parameter that has been universally used in formal semantics since Montague (1974) to classify adjectives is their **entailment pattern**. Montague himself claimed (based, according to him, on work by Kamp and Parsons unpublished at the time)[9] that "The denotation of an

[8] A joke that runs in formal semantic circles is *The meaning of life is life prime*, for in the formal semantics notation, meanings of words are represented by the word followed by a prime symbol: *life'* represents the meaning of *life*.

[9] The relevant pieces of research were later published as Kamp (1975) and Parsons (1970).

adjective phrase is always a function from properties to properties" (Montague, 1974, p. 211)[10] I shall refer to this definition as **Denotation 1**.

To account for the semantics of nominal modification, he defines the *intersection function* in (3.14):

(3.14) [Function] H such that for some property P, H(Q) is, for every property Q, the property possessed by a given individual with respect to a given possible world if and only if the individual possesses both P and Q with respect to that possible world.

This function achieves a result analogous to simple property intersection: given a property such as *red*, and another property such as *ball*, the application of the intersection function will result in a predicate applicable to entities having the property *red* and the property *ball*.

Why does he posit an intersection *function* instead of using simple property intersection? One of the main reasons is that there are adjectives which do not behave like *red*. Montague first cites adjectives such as *big*, which are clearly not intersective in the sense that a big flea is not necessarily a big entity (while a red flea is certainly a red entity). To account for these cases, Montague proposes a weaker version of **Denotation 1**, namely **Denotation 2**: "the denotation of an adjective is a function that always assigns to a property one of its subproperties".[11]

But again, there are some adjectives which do not adjust even to that weaker assumption: *false friend*, *reputed millionnaire*, *ostensible ally*, *possible president*, *alleged intruder* (Montague's examples, p. 211). The solution Montague finds more suited to his purposes is to resort to the general intersection function presented in (3.14), together with some meaning postulates for adjectives. In the following, A stands for an adjective and N for a noun (Montague, 1974, p. 212).

(3.15) every AN is a N

(3.16) an AN is a N such that that N is A

(3.17) every N such that that N is A is a AN

The core distinctions that Montague made have been used throughout subsequent research on adjectives in formal semantics. This work has focused on the two first postulates, establishing a classification as follows:[12]

Intersective or absolute or restrictive adjectives: Those like *red* or *married*, that can be described in terms of intersection and adjust to the three postulates: a red ball is something that is a ball and is red.

[10] A property is the denotation of a unary predicate. A noun such as *table* denotes a property, and a noun with a modifier, as in "big table", also denotes a property. This is why the intersection function takes as input a property and outputs a property.

[11] Note to that respect that a big flea is a flea.

[12] Note that Montague did not name the classes. I will use the first term for each class; see Hamann (1991) for an overview of the treatment of adjective semantics in formal semantics and an explanation of the other terms.

Subsective or relative adjectives: Those like *big*, that do not adjust to postulate (3.16). As mentioned above, a big flea is not something that is big. Hoewver, a big flea is a flea, so that postulate (3.15) is met.[13]

Intensional or nonpredicative adjectives: Those like *alleged*, that do not adjust to any of the three postulates, so that their denotation is a quite unconstrained property-to-property function. An alleged intruder is not necessarily an intruder (postulate (3.15)), and it is even almost ungrammatical to say something like *#That intruder is alleged* (postulate ((3.16)). In fact, a striking property of intensional adjectives is that they cannot be used as predicates (hence the term *nonpredicative*) in many languages, including English and Catalan.

The rest of the Section is devoted to discussing some problems with this classification.

3.3.2.1 Intersective vs. subsective adjectives

The distinction between intersective and subsective adjectives is very difficult to make and to use for classificatory purposes. In fact, Lahav (1989) argues that there are no truly intersective adjectives because "their semantic contribution to the meaning of the whole in which they are embedded varies non-systematically across linguistic contexts" (Lahav, 1989, 261). [14] The variation in meaning, according to Lahav, depends on the head noun in each context.

For instance, he argues, the truth conditions for *X is red* are different depending on whether X is a bird or a table (in the first, but not in the second case, it has to be its natural colour), an apple or a watermelon (an apple has to be red outside, a watermelon inside), a house or a car (a red house only needs the outer walls to be red, but a car needs the whole surface, including doors and roof). Following this line of reasoning, not only the parts that are red or the origin of the colour, but also the shade is different depending on the head noun: a red face does not have the same colour as a red car.

According to Lahav (1989, 266), thus, "the applicability conditions of an adjective are a patchwork of merely related, and not uniform, conditions". These conditions depend fundamentally on the head noun, similarly to size subsective adjectives.

Difficulties in determining applicability conditions grow if the head noun does not denote a tangible entity. Take for instance adjective *abatut*, which in one of its senses can be translated as "depressed". It can be argued that *un professor abatut* "a depressed teacher" is both a professor and depressed, so that it seems that the adjective is intersective. However, the nouns that appear with 'abatut' in the CTILC corpus are mostly abstract nouns: *aspecte* ('aspect'), *estil* ('style'), *ànim* ('mood'), *actitud* ('attitude'), *cor* ('heart'). It really is difficult to decide whether an *aspecte abatut* is an *aspecte* such that this *aspecte* is *abatut*, that is, whether the adjective meets meaning postulate (3.16). [15]

The treatment of adjective semantics in terms of entailment patterns is problematic in that it is not clear how to use postulates (3.16) and (3.17) as a criterium for classification, and more so if we consider abstract head nouns. In addition, it is not clear, from a semantic point of view,

[13] Size adjectives were given an intersective analysis in Kamp (1975); see Section 3.3.2.1.

[14] This characteristic is also termed *syncategorematicity*; see Picallo (2002, fn. 18, p. 1653).

[15] The difficulty was attested in a real classification experiment with human judges, the first Gold Standard experiment reported in Section 4.1 below.

that it is possible to draw the line between intersective and subsective adjectives. Note to this respect that intersective and subsective adjectives are grouped together in the class of qualitative adjectives in the semantic classifications established in descriptive grammar (Section3.2.1).

Entailment patterns are also problematic in another respect: they leave relevant semantic differences among adjectives uncovered. As Louise McNally (p.c.) put it, “subsectivity comes in many flavours”, and the source of the subsective behaviour is presumably different for the different flavours. Kamp (1975) initiated a line of research aimed at reanalysing as intersective apparently subsective adjectives. He analysed scalar adjectives such as *big*. In his account, *big* is intersective but bears a context-dependent index determining the standard of comparison that is being used in each occurence. This analysis corresponds to the intuition that a big flea is big *for a flea*, that is, in relation to the standard size of fleas.

At least two further kinds of apparently subsective adjectives have been recently given an intersective analysis. The first kind of adjective can be termed *event-modifying*, and it has been independently modelled as intersective in two theoretical frameworks: formal semantics (Larson, 1998) and Generative Lexicon (Pustejovsky, 1995; Bouillon, 1999). Both analyses have many points in common, mostly the fact that they “broaden” the lexical structure of the noun to explain the subsective behaviour of these adjectives. I explain the analysis in Larson (1998) and then comment on Bouillon (1999).

Larson (1998) discusses the two readings of a sentence like (3.18).

(3.18) Olga is a beautiful dancer

In one (intersective) reading, Olga is both beautiful and a dancer. In the other (subsective) reading, Olga is a dancer, but what is beautiful is the way she dances, independently of whether Olga is beautiful or not.

Intuitively, the ambiguity involves more what *beautiful* is modifying (the individual Olga or her dancing) than the lexical semantics of the adjective. That is, it involves the semantics of the noun, rather than that of the adjective. To account for this intuition, Larson argues that the noun *dancer* has an event argument corresponding to the event of dancing, in addition to the individual argument for the person to which it is applied. An adjective such as *beautiful* can modify either of the arguments, giving rise to the ambiguity in example (3.18). Schema (3.19) reproduces schema (13) in Larson (1998), which sketches the two analyses of the sentence. [16]

(3.19) a. Qe[dancing(e,olga) ... beautiful(olga,C)] (“Olga is beautiful”)

b. Qe[dancing(e,olga) ... beautiful(e,C)] (“Dancing is beautiful”)

This analysis explains the intersective properties of the adjective while at the same time explaining its subsective behaviour in terms of entailment patterns.

Bouillon (1999) discusses a similar case, *vieux maire* (‘old mayor’), that has two interpretations: one in which the individual described as *maire* ‘mayor’ is aged, and another one in which the mayor job is “of long standing”, that is, the individual, though not necessarily old,

[16] *Q* is a quantifier that binds *e*. *C* is the reference class to determine whether Olga is beautiful or not. The analysis becomes more concrete in the paper, but the schema in (3.19) is sufficiently illustrative for the purposes of the present discussion.

has been mayor for a long time. The analysis that Bouillon proposes is similar in spirit to the Larsonian analysis, in that she posits several arguments for the noun *maire*. Of interest here are the individual argument and an event argument for the process of leading a city. When combining it with adjective *vieux*, which selects for those two kinds of argument,[17] the adjective can modify either of the arguments, yielding the two interpretations. Bouillon's analysis goes one step further than Larson's, in positing event arguments for a noun, *maire* ('mayor'), which is not directly event-denoting.

The second kind of apparently subsective adjectives that have been given an intersective analysis are relational adjectives (McNally and Boleda, 2004). Their main properties have been reviewed in Section 3.2.2. They have not received much attention in formal semantics or generative linguistics (though see Levi (1978), Bosque and Picallo (1996) and Fradin and Kerleroux (2002)) From the point of view of formal semantics, they can be treated as subsective because of their entailment patterns, as depicted in example (3.20) (McNally and Boleda, 2004, ex. 1, p. 179).

(3.20) a. El Martí és arquitecte tècnic.
'Martí is a technical architect'

b. $\models$ El Martí és arquitecte.

c. $\not\models$ #El Martí és tècnic.

(3.20a) entails that Martí is an architect (meaning postulate (3.15)) but not that he is technical (meaning postulate (3.16)); indeed, *tècnic* sounds rather anomalous when applied to *Martí*. *Tècnic* is not a prototypical intersective adjective.

McNally and Boleda (2004) suggest for relational adjectives an analysis along the lines of Larson's analysis for event-modifying adjectives. Under their analysis, relational adjectives are properties of *kinds*, not of individuals. [18] They posit an implicit kind argument for all common nouns, so that "objects realize the kinds of things that nouns describe" (McNally and Boleda (2004, p. 188)). Relational adjectives modify the kind argument, not the individual argument, which accounts for the entailment patterns shown in (3.20). [19]

[17] In fact, Bouillon proposes that it selects for a single *dotted type* argument, a complex type composed of an individual and an event. Dotted types are one of the technical devices introduced in the Generative Lexicon framework (Pustejovsky, 1995). They account for the fact that some words denote two different kinds of entities "at the same time": a book is something that is both a physical entity and a piece of information. The two kinds of entities interact, as in *Eve believes the book* (Bouillon, 1999, ex. (9a), p. 152), in which what Eve believes is the information contained in a physical object.

[18] *Kinds* "are modeled as special types of individuals" (Krifka et al., 1995, p. 64). Some NPs can be said to provide a "*reference to a kind* –a genus– as exemplified in [3.21]. The underlined noun phrases (NPs) in [3.21] do not denote or designate some particular potato or group of potatoes, but rather the kind Potato (*Solanum tuberosum*) itself. [An NP of this type] does not refer to an "ordinary" individual or object, but instead refers to a kind." (Krifka et al., 1995, p. 2; examples in (3.21) reproduce their examples in (1))

(3.21) a. The potato was first cultivated in South America.

b. Potatoes were introduced into Ireland by the mid of the 17th century.

c. The Irish economy became dependent upon the potato.

See Carlson (1977) and Krifka et al. (1995) for further information about kinds.

[19] The intuition is present in Bosque and Picallo (1996) and Picallo (2002). For instance, (Picallo, 2002, 1667) asserts that relational adjectives "usually refer to generic objects with respect to which the reference of the nominal head is delimited" ("Solen fer referència a objectes genèrics respecte als quals es delimita la referència del nom que és el nucli del sintagma." My translation (GBT).). However, neither work develops this intuition into a specific semantic analysis.

This analysis also accounts for a number of properties of relational adjectives. The most prominent one is the fact that they can appear in predicative positions under very restricted circumstances, namely, when the subject refers to a kind rather than an individual. Note that while (3.20c) is anomalous, (3.22) is perfectly natural.

(3.22) La tuberculosi pot ser pulmonar.
'Tuberculosis can be pulmonary'

In (3.22), *La tuberculosi* does not denote an entity, but a kind, and *pulmonar* restricts its denotation to one of its subkinds. McNally and Boleda's analysis thus captures the classificatory flavour of relational adjectives, noted by many researchers: if they are properties of kinds, their main function will be to establish subkinds, that is, to further classify entities. How the analysis technically works is not relevant here (see McNally and Boleda (2004)).

Event-modifying and relational adjectives show that a semantics for adjectives solely based on inferential patterns obscures crucial differences in the types of subsectivity found in adjectives of natural languages. To account for these differences, a richer structuring of the semantics of the nouns is needed. Also, compositional devices that allow different kinds of composition of the meaning of the adjective and its (nominal or other) argument have to be developed.

The discussion of adjective *red* shows that, even for prototypical intersective adjectives, the postulates do not strictly hold. Although some intersective adjectives are sharper than *red* (e.g., 'solter' *bachelor*), what we learn is that the line between subsective and intersective adjectives is very difficult to draw, if at all possible. Meaning postulates 3.15 and 3.16 can not easily be used as criteria for classification. A further argument for blurring the distinction is that no syntactic class properties distinguish subsective from intersective adjectives. None of parameters discussed in Section 3.6.1 yield consistent differences between intersective and subsective adjectives. If neither entailment patterns nor syntactic behaviour provides a reliable means to distinguish between subsective and intersective adjectives, the distinction has to be abandoned for the present purposes. The two kinds of adjectives will be collectively referred to as intersective or qualitative.

3.3.2.2 Intensional adjectives

The same examples for intensional adjectives are given over and over again: *former* and *alleged* (as in most textbooks on formal semantics, such as Dowty et al. (1981), Lappin (1996), Heim and Kratzer (1998), Chierchia and McConnell Ginet (2000)). Montague further mentions *false*, *reputed*, *ostensible*, and *possible*. It is indeed a short list compared to the 17,281 adjective lemmata in GLiCom's computational dictionary.

From a truth-conditional point of view, adjectives such as *former* or *alleged* are indeed remarkable, and their behaviour should be analysed and explained. However, a criticism can be made, from both a theoretical and a practical point of view, of the fact that the main semantic division within the adjective category separates (and thus characterises) a very small percentage of the category from the rest. This is so because the vast majority of adjectives do what adjectives are supposed to do: specify a property of its argument. It may be a very typical property, such as *red* or *large*, or a not so typical property, such as *autònom* 'autonomous' or *subaltern* 'minor, auxiliary' (for personnel); but it is a property, and thus most adjectives adjust at least to the inferential behaviour encoded in meaning postulate 3.15.

In addition, intensionality, like subsectivity, comes in many flavours. *Former* has to do with contrasting times of utterance and reference. *Alleged*, *reputed* and *ostensible* appeal to the attitude of the speaker. *Possible* raises issues of modality. *False* is a so-called "privative" adjective (Partee, 2001), that is, it belongs to the reduced subset of nonpredicatives which entail the negation of the nominal property. The reasons for each of these kinds of adjectives not adjusting to any of the meaning postulates presumably have to be traced back to different sources.

In fact, we find in the literature proposals to reanalyse some intensional adjectives as intersective or subsective. Partee (2001) analyses privative adjectives such as *fake*, *counterfeit*, or *fictious*. Recall from Section 3.3.1 that intensional adjectives can not be used as syntactic predicates. Privative adjectives have the remarkable property that, although they are arguably nonpredicative in meaning (a fake gun is not a gun), they can be used as predicates, as in example (3.23) (Partee, 2001, ex. (10a)).

(3.23) Is that gun real or fake?

Partee also argues that the very interpretability of example (3.23) is a puzzle that remains unexplained with an intensional analysis of *fake*. She further uses data from NP-split in Polnish to argue that syntactically, privative adjectives are grouped with subsective adjectives, not with intensional adjectives. To explain these pieces of data, she proposes that *fake* expands (*coerces*) the denotation of *gun* to include both fake and real guns. Once the coercion has taken place, nothing prevents the adjective from modifying the noun in an intersective way.

Partee claims that the default denotation of *gun* only includes real guns, "as is evident when one asks how many guns the law permits each person to own, for instance" (p. 5). However, the default denotation is highly context-dependent: in a toys store, presumably the manager would ask "how many guns are there left?", not "how many fake guns are there left?". Thus, the coercion can be triggered by other, contextual factors, apart from linguistic cues. Partee acknowledges that "the extension of nouns is quite 'adjustable' "; the toy store is presumably one of the cases where the extension is adjusted.

Modal adjectives such as *possible* bear some resemblance to privative adjectives. They have an intensional meaning: a possible president is not necessarily a president (postulate (3.15)). They have the characteristic that explicitly introduce possible worlds. However, they can be used as predicates. The examples that follow are attested in the CTILC corpus.

(3.24) a. No és possible fer un casament laic?
Not is possible do a marriage lay?

'Isn't it possible to have a lay marriage?'

b. És possible que necessitis el cotxe
Is possible that you-need the car

'You may need the car'

c. La pau no és possible
the peace not is possible

'Peace is not possible'

d. Quants? L'avaluació del nombre, si fos possible, que no ho és, ens
How-many? The-evaluation of-the number, if were possible, that not it is, us
ajudaria a fixar els termes del procés
would-help to fix the terms of-the process

'How many? The evaluation of its number, if it were possible, which it is not, would help us fix the terms of the process'

e. . . . serveixen, si més no, per a mostrar quins mitjans diversificats de transició
. . . serve, if more not, for to show which means diversified of transition
resulten possibles
result possible

. . . '[they] serve, at least, to show which diversified transition means turn out to be possible'

Because *possible* mainly modifies proposition-denoting constituents, the most typical environment for predicative *possible* is the case where the subject is a clause, either an infinitival clause (example (3.24a)) or a *that*-clause (example (3.24b)). However, cases with noun phrase subjects are also possible, as in examples (3.24c) and (3.24d). Predicative uses in other environments than the copular sentence are also attested (example (3.24e)). The same applies to other modal adjectives, such as necessari or *necessari* ('necessary') or *contingent* ('contingent'). I will not pursue an intersective analysis of modal adjectives. I just want to remark that their syntactic behaviour remains unexplained if we assign them to the intensional class, typically nonpredicative. Further syntactic arguments regarding semantic classification are given in Section 3.6.1.

3.3.3 Summary

In this Section, we have argued that the treatment of adjective meaning in formal semantics is not adequate for our purposes. Entailment patterns raise interesting questions from a logical point of view, and uncover behaviours (such as those of intensional adjectives) that have to be analysed and explained. However, if used on their own, entailment patterns have three main problems.

On the one hand, they generate a distinction that is very difficult to make, if at all possible: intersective vs. subsective adjectives. Clear intuitions can only be obtained for very specific adjectives that modify concrete objects or humans. Even in these cases, the meaning of an adjective changes depending on the noun it modifies and, more generally, on context[20]. Thus, if taken literally, meaning postulate (3.16) does not apply to any adjective.

Lahav (1989) pushes the argument as far as to deny the possibility of compositional semantics, that is, of establishing the meaning of a larger utterance from the meaning of its parts. If the denotation of *red* varies in an unpredictable fashion depending on what it modifies, then the task of compositional semantics indeed faces a dead end. However, the shifts in meaning do not seem to be truly unpredictable, so that the position can be weakened to claim that there are no truly intersective adjectives.

The second problem about using entailment patterns as classification criteria is that they leave major kinds of semantic differences uncovered. As we have seen in the discussion of event-

[20] This characteristic has been widely acknowledged and referred to as *plasticity*, *flexibility*, *vagueness*, *syncategorematiciy*, etc.

modifying and relational adjectives, as well as the discussion of intensional adjectives, the reasons for each (type of) adjective not meeting the postulates are different in each case. If the members within each class differ both semantically and syntactically in relevant respects, the classification loses predictive and explicative power.

The third problem is that one of the classes, intensional adjectives, is very small. This is a concern particularly for practical NLP purposes, such as developing a large-scale lexicon, because a distinction that only affects a very small number of lexemes will not be very useful unless these items are extremely frequent (as is the case with closed-class part of speeches, for instance). In Machine Learning terms, a class that is very small will be very difficult to successfully identify. From a theoretical point of view, however, a very small adjective semantic class again raises concerns of predictability and explicability. The intensional class was part of the classification in early stages of the research presented here. It was later abandoned for the reasons explained here and for consistency reasons (the parameter that was used for the classification was the ontological type of the denotation, which does not coincide with entailment patterns). This process will be explained in Chapter 5.

3.4 Adjectives in ontological semantics

3.4.1 General framework

One of the alternatives for meaning postulates to account for lexical semantics was Katz's decompositional semantics (Katz and Fodor, 1963; Katz, 1972). [21] Katz represents senses of words as structured entities called *semantic markers*. They are trees with labeled nodes, whose structure mirrors the structure of the sense. The labels identify the sense's *conceptual components*. For instance, for the verb 'chase', some of the labels are ACTIVITY, PHYSICAL, MOVEMENT, PURPOSE, CATCHING, and the tree structures them so that, e.g., MOVEMENT is a daughter of PHYSICAL, which is a daughter of ACTIVITY (see Marconi (1997, p. 21)).

The idea in decompositional semantics is to describe the internal structure of lexical meanings with a number as small as possible of semantic markers and their relationships, so that in constructing more complex meanings the possible inferences are explicitly represented. The inference "Paul moved" from "Paul chased Bill" is explicitly represented in stating that the ACTIVITY denoted by 'chase' is of a PHYSICAL MOVEMENT type.

According to Nirenburg and Raskin (2004, p. 99), Katz and Fodor (1963) showed that "the general lexicon could be represented using a limited number of semantic features only if one agreed to an incomplete analysis of word meaning." There is always a part of the meaning of the word, a *semantic distinguisher* in Katz and Fodor's terms, left undescribed after a componential analysis. In fact, Nirenburg and Raskin argue, "if the number of primitives is kept small, descriptions tend to become complex combinations of the primitives that are hard to interpret and use." This argument leads Nirenburg and Raskin to defend a rich system of semantic primitives with complex interrelations, namely, an ontology. The definition and role of ontologies have been analysed for a long time within Philosophy, and further developed within artificial intelligence (Gruber, 1993).

The integration of an ontology within an articulated theory of lexical and compositional seman-

[21] "Whether [Katz's theory is] equivalent to meaning-postulate theory is a matter of controversy" (Marconi, 1997, p. 20).

tics lead to Ontological Semantics, a theoretical framework developed by Sergei Nirenburg and Viktor Raskin (Nirenburg and Raskin, 2004). One of its main premises is that "text meaning involves both linguistic and world knowledge; is necessary for advanced computational-linguistic applications; is extractable and formally representable." (Raskin and Nirenburg, 1998, 136).

One of the main motivations for the theory is precisely the machine tractability of the representations, so that meanings can be modeled through the computer. In NLP, where the agent of the analysis (the computer) has no access to the world, it becomes even clearer that world knowledge has to be encoded to provide the framework for inferencing that is necessay in many languages understanding tasks.

In Ontological Semantics, there are three so-called static knowledge sources: the ontology, the fact repository and the lexicon (Nirenburg and Raskin, 2004, Chapter 7). Only the lexicon is language-dependent.

The ontology "contains the definitions of concepts that are understood as corresponding to classes of things and events in the world" (Nirenburg and Raskin, 2004, p. 191). [22] The fact repository contains entries for instances of these concepts (e.g. Sydney Olimpics for the concept SPORTS-EVENT, or Barcelona for the concept CITY). The lexicon contains the linguistic information (not only semantic, but also morphological and syntactic) concerning lexical items. The semantic information in the lexicon specifies the concepts, properties, or properties of concepts included in the Ontology that account for the meaning of a word. [23]

Most semantic theories presuppose a tight relationship between syntax and semantics, so that the output of the syntactic component serves as the input for the semantic component (Thomason (1974), Katz and Fodor (1963)). Ontological semantics does not commit to this view; instead, it claims a high independence between these levels. This aspect is not in accordance with the work presented here, which heavily draws on the syntax-semantics interface, as will be explained in Section 3.6.

Lewis (1972) criticises Katz's decompositional semantics. He points out the failure of its representational system, which he called "Markerese", to relate language to the extralinguistic reality. We end up with the same problem that meaning postulates had, namely, that defining a language in terms of another language only moves the semantic burden from the first language to the second one. Remember the learning-Chinese-with-a-Chinese-dictionary paradox stated in Marconi (1997, p. 19) and cited in Section 3.3.1.

The same criticism can be made of any representation of meaning, however rich and complex, and it has been repeatedly noted. Because we do not know how the mind represents meanings,

[22] There has been and continues to be much debate in philosophy as to what kind of reality ontologies encode (e.g., whether properties "exist" independently of individuals, as in the realist/naturalist debate). To this respect, Nirenburg and Raskin (2004, p. 135) state:

> What ontological semantics aims to reflect is the use of concepts by humans as they see them, introspectively and speculatively; and people do talk about properties, fictional entities (unicorns or Sherlock Holmes), and abstract entities as existing. For us, however, the decision to include the abstract and fictional entities in the ontology is not motivated by the fact that these entities can be referred to in a natural language. Rather, we believe that languages can refer to them precisely because people have these concepts in their universe.

[23] The lexicon includes an onomasticon, or lexicon of names, that is directly linked with elements of the fact repository, rather than the ontology. It corresponds to the view that proper names denote particular individuals, that are not placed in the ontology.

and meanings cannot be directly manipulated, any representation of a lexical meaning is just a translation into another language. The criticism applies, therefore, to ontological semantics.

However, even if an ontology is uninterpreted, it is a model of the world. We adhere here to the view in Nirenburg and Raskin (2004) that it is necessary from a theoretical and practical point of view to distinguish between extra-linguistic and linguistic concepts to account for the semantics of words, and to encode the extra-linguistic knowledge in an explicit fashion. In fact, most (if not all) semantic theories dealing with lexical semantics use primitives in an explicit or implicit way (as in the Lexical Conceptual Paradigm by Jackendoff (1990) or the Generative Lexicon by Pustejovsky (1995)).

Even in formal semantics, the famous prime notation (see footnote 8, page 30) could be viewed as a way to mark as a primitive every lexical item present in a particular model. Defining an ontology allows for an explicit definition and characterisation of the semantic primitives used, and a clear delimitation of world knowledge and linguistic knowledge.

3.4.2 Adjective classes

In Ontological Semantics, "the most crucial taxonomic distinction within the lexical category of adjectives", as Raskin and Nirenburg (1998, p. 167) put it, "is the ontology-based distinctions among

- scalar adjectives, whose meanings are based on property ontological concepts;
- denominal adjectives, whose meanings are based on object ontological concepts; and
- deverbal adjectives, whose meanings are based on process ontological concepts."[24]

These distinctions affect the components and structure of their lexical entries in addition to the ontological type of their denotations. For scalar adjectives, it is necessary to identify the scale affected (for instance, AGE for adjective *old*). If the scalar is gradable, a numerical value for the scale will be coded. If it is not gradable, a literal value corresponding to the relevant ontological property will be specified (for instance, MAGENTA for the COLOR attribute of the adjective *magenta*).

For object- or process-based adjectives, their semantics contains a link to the relevant ontological entity, together with additional semantic information (for deverbal adjectives, for instance, the semantic role of the nominal argument is encoded). The details of the ontological semantic representation as instantiated in an NLP resource are provided in Section 3.5.3.

The classification I propose (Section 3.6.3) is very similar to this taxonomy, with some caveats that will arise in subsequent discussion.

Raskin and Nirenburg, p. 177 distinguish further subclasses within scalar adjectives, namely, numerical scale (of the SIZE type), literal scale (of the COLOUR type), attitude-based (adjectives such as *good, superb, important*), and member adjectives (including privative adjectives such as *fake*, but also other adjectives such as *authentic, similar, nominal*). The entries of the two

[24] Although the authors refer to the object- and process-based classes in terms of their typical morphological type (denominal or deverbal), the definition is semantic, as they allow nonderived adjectives to be represented the same way (e.g., *eager* related to *want*).

last subclasses, although of a scalar type, are more complex, including for example intersection of properties for member adjectives. The details are not relevant for the present purposes.

3.4.3 Summary

The kinds of problems that motivated the Ontological Semantics framework have many points in common with those that motivate the research presented here. It is not surprising that it turns out to be the most fruitful theoretical proposal for the purposes of this PhD.

The main points in common are the following. The classification I aim at should cover a large portion of the lexicon (coverage is an issue), and should account for the meaning of ordinary adjectives, defined as all adjectives that occur with a some pre-determined frequency. This represents "a shift from the usual focus of theoretical linguistics on exceptional and borderline phenomena ... to a description of ordinary cases" (Raskin and Nirenburg, 1998, 136). Relatedly, the lexical items to be analysed are obtained from corpora, not from a theoretically biased selection. Both research in ontological semantics and the resarch presented here aim at machine tractability of representation, so that generalisation is driven by machine tractability. Finally, in both cases the goal is to automate acquisition of semantic information to the largest extent possible.

There are, of course, many differences between the goals and methodologies followed in this thesis and in the ontological semantics framework.[25] Nirenburg and Raskin do not believe that the syntax-semantics interface can be meaningfully exploited for meaning acquisition and characterisation. This is the most important difference with the research presented here, which heavily relies on the morphology-semantics and syntax-semantics interface. However, Nirenburg and Raskin (2004, p. 123) admit that there are "grammaticalized semantic distinctions", or "semantic phenomena that have overt morphological or syntactic realizations" They argue: "it seems obvious that there are more lexical meanings than syntactic distinctions, orders of magnitude more. That means that syntactic distinctions can at best define classes of lexical meaning ... rather coarse-grained taxonomies of meanings in terms of a rather small set of features" (Nirenburg and Raskin, 2004, p. 124) This is exactly what I will pursue: adjective classes in my definition are classes of lexical meaning obtained from quite a small number of features, that can serve as sketches for lexical semantic entries for subsequent manipulation or further automated acquisition.

3.5 Adjective semantics in NLP

3.5.1 WordNet

WordNet (Fellbaum, 1998b) is currently the most widely used lexical semantics resource in NLP. It is based on a relational view of the lexicon, and thus organises all content words (nouns, verbs, adjectives, adverbs) according to one or more paradigmatic lexical relationships. Synonymy is the basis for the organisation of the whole lexicon, and it gives name to the unit of analysis in WordNet: the *synset*, or set of synonyms. Furthermore, nouns are organised via hyponymy (Miller, 1998a) and verbs via troponymy (Fellbaum, 1998a).

[25] The most obvious one is the development of an ontology in parallel to lexicon, which is too formidable a task for it to be feasible within a PhD.

As for adjectives, antonymy is the basic organising paradigmatic relationship of the category (Miller, 1998b). WordNet distinguishes between two kinds of adjectives, *descriptive* and *relational*. Descriptive adjectives are characterised according to Miller (1998b) as denoting attributes and can be organised in terms of antonymy (they correspond to qualitative, intersective/subsective, and scalar adjectives as termed in Sections 3.2, 3.3, and 3.4, respectively). Relational adjectives, in general, cannot be organised this way. Within descriptive adjectives a subclass is further distinguished, that of *participial adjectives*, which includes gerunds and participles.

The organisation of adjectives through antonymy is problematic at least in two ways, both of them mentioned in Miller (1998b). The first problem is that only a subclass of adjectives, those termed *descriptive adjectives* in WordNet, contains antonym-bearing adjectives. Relational adjectives (which in WordNet are directly identified with denominal adjectives) do not in general have antonyms (*apolític* 'apolitical' is not the antonym of *polític* 'political'; see Demonte (1999)). For this kind of adjectives, two different solutions are adopted. If a suitable antonym can be found (antonym in a broad sense; in Miller (1998b, 60), *physical* and *mental* are considered antonyms), they are treated in the same way as descriptive adjectives. Otherwise, they are listed on their own, with a pointer to the deriving noun.

As for participial adjectives, subclass of descriptive adjectives, they also receive a hybrid treatment. Those that can be accomodated to organisation through antonymy are treated as descriptive adjectives. An example of the author is *laughing - unhappy*, which are indirect antonyms (Miller, 1998b, 58). We will review this notion below. Those that cannot be accomodated, like *obliging* or *elapsed*, are related with the deriving verb through a PRINCIPAL-PART-OF link.

The second problem about the treatment of adjectives in WordNet is that, even within the descriptive class of adjectives, only a few actually have lexical antonyms. Miller (1998b, 49) argues that "Antonymy, like synonymy, is a semantic relation between word forms", that is, it is a lexical, not a conceptual relationship. For instance, even if *large* and *big* are very similar concepts, they have different antonyms: *small* is the proper antonym for *large,* but not for *big* (which has as antonym *little*). There are many adjectives that do not have a lexical antonym (*direct* in the WordNet terminology), and that is the reason why the concept *indirect antonym* was established in WordNet.

Adjectives that do not have a direct antonym are grouped according to similarity of meaning around a *focal adjective*, which does have a direct antonym.[26] The focal adjective's antonym is the indirect antonym of a whole group of adjectives. For example, *fast* and *slow* are direct antonyms, and *swift* and *slow* are indirect antonyms, through the semantic similarity detected between *swift* and *fast*.

In parallel to that organisation, and when an adjective expresses values of an attribute lexicalised through a noun (e.g. *deep, shallow* and *depth*; see Miller (1998b, 54)), a link between adjective and noun is established, so that the nominal and adjectival lexical hierarchies are connected.

The main problem of the WordNet approach to adjective semantics is that it attempts at classifying a whole part of speech through a lexical relationship that only applies to a small subset of this class. In WordNet 1.5 there are 16,428 adjectival *synsets*, and only 3,464 (slightly over 20%) are organised through antonymy, including direct and indirect antonymy. Assuming that

[26] The notion of similarity of meaning is, of course, problematic. Miller (1998b, 50): "The term *similar* as used here typically indicates a kind of specialization; that is to say, the class of nouns that can be modified by *ponderous*, for example, is included in -is smaller than- the class of nouns that can be modified by *heavy*".

there are about ten indirect antonyms per direct antonym pair (five per antonym), we can estimate that antonymy proper can be applied to less than 3% of the adjectives in the lexicon only. Most of the adjectives are simply linked to a noun or verb.

3.5.2 SIMPLE

The goal of the SIMPLE project was to create a standard for the representation of lexical semantic information to be included in computational lexica. Peters and Peters (2000) describe the theoretical basis for the treatment of adjectives in SIMPLE.

The basic distinction in SIMPLE is "the predicative type distinction between extensional and intensional adjectives", that is, the difference between adjectives like *crazy, soft* and adjectives like *former, present, necessary* (Peters and Peters, 2000, 7). Within extensional adjectives, they distinguish between intersective and subsective adjectives.

This proposal is based on entailment patterns as explained in Section 3.3, because Peters and Peters (2000) consider that information about possible entailments can be very useful in NLP. As an example, they argue that if a system knows that *American* is intersective and *former* intensional, it can deduce that a person referred to by *American president* is president in reference time, while if the phrase is *former president* that person is not president in reference time. In addition, according to the authors, intersective adjectives that refer to the same meaning component usually do not modify expressions that refer to the same entity, while subsective adjectives can do so. For instance, *red car* and *blue vehicle* almost surely refer to different entities, and that is not the case for *large mouse* and *small creature*.

Within each of the classes defined according to entailment patterns, the proposal distinguishes further subclasses on the basis of denotational properties, in the fashion of Dixon (1982). A summary of this "Adjective Ontology" (Peters and Peters, 2000) is listed in what follows:

1. Intensional adjectives

   (a) Temporal: *former president, present situation*

   (b) Modal: *certain victory, necessary ingredient, potential winner*

   (c) Emotive: *poor man*

   (d) Manner: *beautiful dancer*

   (e) Object-related: *criminal lawyer*

   (f) Emphasizer: *outright lie*

2. Extensional adjectives

   (a) Psychological property: *crazy thoughts*

   (b) Social property: *catholic priest*

   (c) Physical property: *soft skin*

   (d) Temporal property: *sudden outburst*

   (e) Intensifying property: *heavy rain*

   (f) Relational property: *similar shape*

The problem with the basic division between intensional and extensional adjectives is the same as reviewed in Section 3.3: there are very few intensional adjectives, and thus the basic distinction separates a very small group from a very large one with characteristics and internal differences that are left unexplained. In addition, many of the examples cited by Peters and Peters (2000) as belonging to the intensional class are problematic.

Beautiful is clearly intersective in sentences like *Olga is beautiful*. Regarding the eventive modification (*dances beautifully*), its behaviour is subsective, not intensional. In addition, Larson (1998) provided an intersective analysis even of this kind of modification, as explained in Section 3.3.2. The same applies to relational adjectives. They are coded as a subclass of intensional adjectives (called *Object-related*). However, the entailment pattern they follow corresponds to subsective, not to intensional adjectives:

(3.25) a. John is a criminal lawyer $\models$ John is a lawyer

b. John is a criminal lawyer $\not\models$ #John is criminal

These adjectives have also been given an intersective analysis (McNally and Boleda, 2004, see Section 3.3.2). It is also not clear what distinguishes 'object-related' from 'social' adjectives, and yet they are placed in two separate classes.

Finally, not all modal adjectives exhibit an intensional entailment pattern, as can be seen in the sentences in 3.26. Modal adjectives are a peculiar subclass within intensional adjectives, as has been discussed in Section 3.3.2.2.

(3.26) a. Patience is a necessary ingredient [for the resolution of X] $\models$ Patience is necessary

b. John is a potential winner of this race $\not\models$ #John is potential

(Peters and Peters, 2000) acknowledge that "intensional adjectives do not form a semantically homogeneous class", but the fact is that they are also neither morphologically nor syntactically homogeneous.

As for the ontological distinctions, they suffer from the usual problems in any denotational classification: incompleteness and fuzziness. The suggested categories cover a wide range of the possible meanings of adjectives, but they do not cover for instance adjectives such as *fake*. In addition, the categories seem to be fuzzy: why is *sudden* in *Temporal property*, and not, for instance, in *Manner*? Or why is there no *Manner property*?

These are problems that probably arise in any ontological account, and should not be an obstacle to continue trying to define an adequate ontology for lexical meanings. What is perhaps more problematic is to mix a formal semantic criterion with an ontological criterion within the same hierarchy, providing a hybrid kind of resource where the distinctions do not form a consistent set of conditions.

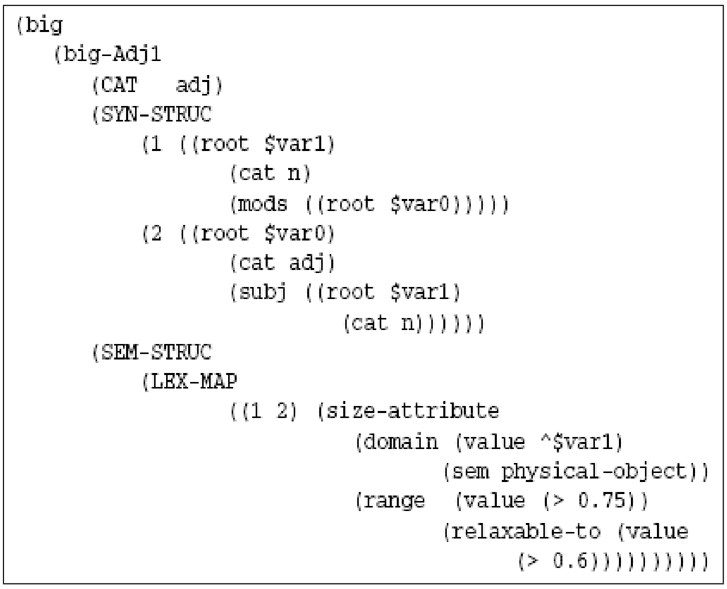
3.5.3 MikroKosmos

The MikroKosmos system (Onyshkevych and Nirenburg, 1995) is one of the implementations of the ontological semantics framework as a system for natural language processing. The treatment of adjectives within MikroKosmos generally corresponds to what is explained in Section

3.4.2 above. In this Section we outline more specific aspects of the implementation of the theory in an actual system, as explained in Raskin and Nirenburg (1996; 1998).

In MikroKosmos, as in other lexical frameworks such as Head-Driven Phrase Structure Grammar, the lexical entries are structured into two main areas, one containing syntactic information and another one containting semantic information. A third area in MikroKosmos specifies the part of speech (CAT) of the word.

For adjectives, the syntactic information is always the same, and covers the noun-modifying and predicative functions of the adjective. In Figure 3.1, which reproduces example (32) of Raskin and Nirenburg (1998, p. 164), it is represented in the SYN-STRUC area. The modified noun is assigned the variable $var1, and the adjective the variable $var0. The sub-area 1 under SYN-STRUC represents the modifying function: the syntactic root is $var1, of category noun (head noun), and its modifier ('mods') is headed by $var0, the adjective. The sub-area 2 under SYN-STRUC represents the predicative function. This time, the syntactic root is $var0, the adjective, and its subject ('subj') is headed by the noun represented by $var1.



```
(big
   (big-Adj1
      (CAT   adj)
      (SYN-STRUC
          (1 ((root $var1)
                 (cat n)
                 (mods ((root $var0)))))
          (2 ((root $var0)
                 (cat adj)
                 (subj ((root $var1)
                            (cat n))))))
      (SEM-STRUC
          (LEX-MAP
                 ((1 2) (size-attribute
                             (domain (value ^$var1)
                                    (sem physical-object))
                             (range  (value (> 0.75))
                                    (relaxable-to (value
                                         (> 0.6))))))))))
```

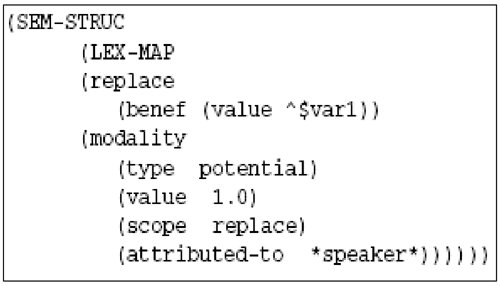
Figure 3.1: *Representation of adjective big in MikroKosmos.*

No discussion is provided of other syntactic environments, such as predicative function with non-copular verbs, as in *I saw her drunk*. The possibility that the modified head is not a noun, but a clause, is also not discussed. In English, in these cases usually the head is a "dummy" *it*-pronoun, as in *It is very important that you come*. In Catalan and other languages, it is possible for the clause to formally act as the subject of the copular sentence, as in (3.27).

(3.27) Que vinguis és molt important per a mi
That come is very important for to me

'It is very important for me that you come'

As for the semantic area, SEM-STRUC, it mainly specifies the mapping to the ontology (LEX-MAP area). *Big*, being a scalar adjective, is mapped onto an *attribute* kind of concept, namely, the *size-attribute*. The *domain* to which it is applied is the meaning of the head noun, formalised by applying a caret, '^', to the nominal variable $var1. Selectional restrictions, such as the noun



```
(SEM-STRUC
      (LEX-MAP
      (replace
         (benef (value ^$var1))
      (modality
         (type  potential)
         (value  1.0)
         (scope  replace)
         (attributed-to  *speaker*)))))
```

Figure 3.2: *Representation of adjective replaceable in MikroKosmos.*

being a *physical object*, are specified under *sem*. These are also pointers to an element in the ontology.

The range of the attribute is specified under *range*. In MikroKosmos, by convention, scalar concepts are assigned numerical values ranging from 0 to 1. *Big* is assigned a value higher than 0.75, that can be lowered to higher than 0.6 (*relaxable-to (value* > 0.6*)*). The *relaxable-to* slot is one of the means to encode *defeasible information* in MikroKosmos, an important matter when building ontologies for natural language. [27]

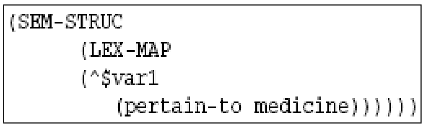
When combining a scalar adjective with a noun, the analysis proceeds "inserting its meaning (a property-value pair) as a slot filler in a frame representing the meaning of the noun which this adjective syntactically modifies" (Raskin and Nirenburg, 1996, 843)

It could be argued that the treatment of scalar adjectives in MikroKosmos (and hence in ontological semantics) is subject to the same criticism that WordNet was subject to. There are not many typical scales such as age, temperature, colour, etc., in the same way that the antonymy relationship only applies to a small subset of adjectives. For instance, Catalan *autònom* ('autonomous'), *íntegre* ('integer'), or *perillós* ('dangerous') cannot be placed in any typical scale. Should an AUTONOMY, INTEGRITY, or DANGER scale be postulated? It is easy to see that, for this kind of approach, it could be necessary to postulate as many scales as adjectives for large portions of the vocabulary. Indeed, "The typology of scales for scalars ... emerges as a major issue in adjective semantics and lexicography" Raskin and Nirenburg (1998, p. 172) .

However, because scalar adjectives are characterised by being related to an attribute, an alternative view is possible. It can be argued that scales form naturally when several adjectives relate to the same attribute, but it is not a necessary condition for adjectives to cluster around scales, as was the case in WordNet. The attribute concepts AUTONOMY, INTEGRITY, and DANGER do make sense, and this is all that is needed for an encoding along the lines of Figure 3.1. In this view, not all adjectives termed *scalar* in Raskin and Nirenburg (1998) are straightforwardly scalar. For this reason, I will avoid the term scalar and use the term *basic* instead (because the most prototypical adjectives belong in this class).

Event-derived or process-related adjectives are mapped onto *event* concepts in the ontology. The idea is that the meaning of an adjective is literally derived from that of the deriving verb. The representation of adjective *replaceable* is depicted in Figure 3.2, which reproduces part of example (62a) of Raskin and Nirenburg (1998, p. 187; CAT and SYN-STRUC areas are omitted for conciseness, as they are identical to the same areas in Figure 3.1). Two pieces of information

[27] The idea is that there is a default or prototypical use of words: for instance, it is safe to assume that birds fly, despite the fact that some birds do not. The information about words in any taxonomy of natural language should be encoded so that default knowledge can be used in the general case, but that it can be overriden when necessary.



```
(SEM-STRUC
      (LEX-MAP
      (^$var1
         (pertain-to medicine))))))
```

Figure 3.3: *Representation of adjective medical in MikroKosmos.*

are supplied in addition to the semantics of the corresponding verb: first, the thematic role (agent, theme, etc.) filled by the head noun. For *replaceable*, the role is beneficiary (‘benef’), and its value is assigned to the meaning of the head noun, $var1. Second, the semantics added by the morphological derivational process. In the case of *replaceable*, or in general for the *-ble* morpheme, the information is “the positive potential attitude”, specified under *modality* in Figure 3.2.

Raskin and Nirenburg (1998, p. 187) note: “There are cases of semantic “suppletivism,” when the entry for an adjective is derived from a semantically immediately related but morhpologically unrelated verb”. They cite examples such as *audible*, from *hear*, or *ablaze*, from *burn*. This suppletivism is not limited to event-related adjectives, but can also be found in object-related adjectives. We will return to this issue in Section 3.6.2.

Object-related adjectives are mapped onto object concepts in the ontology. The semantic subarea of adjective *medical* is depicted in Figure 3.3, which reproduces the SEM-STRUC part of example (64b) of Raskin and Nirenburg (1998, p. 189). The meaning of an adjective is derived from that of the deriving noun. However, for most object-related adjectives, the relationship with the nominal head is not as specific as for the *-ble* adjectives. For these cases MikroKosmos specifies a “the catch-all relation PERTAIN-TO” (Raskin and Nirenburg, 1998, p. 189).

In Figure 3.3, the expression under LEX-MAP expresses that the meaning of the head noun, ^$var1, has a generic PERTAIN-TO relationship to the concept ‘medicine’. More specific relations are defined for adjectives such as *federal* (OWNED-BY *federation*) or *malignant* (HAS-AS-PART *cancer-cell*; see Raskin and Nirenburg (1998, pp. 189-192)).

3.5.4 Summary

In this Section we have reviewed how resources that are specifically designed for NLP purposes encode information related to lexical semantics.

The three resources reviewed have in common the attention devoted to denotational properties: all three distinguish attribute-denoting adjectives from other kinds of adjectives, such as object-related, modal, or event-related.

In fact, the classification in WordNet quite closely resembles the one adopted in MikroKosmos, with descriptive adjectives corresponding to scalar, denominal to object-related, and participials, a subclass of descriptive, to a subset of event-related adjectives. The fact that in WordNet other kinds of deverbal adjectives (with suffixes *-ive*, *-ble*, *-or*, etc.) are not separately treated presumably corresponds to the fact that the semantic relationships they establish with the deriving verb are quite specific. SIMPLE proposes a hybrid classification, with a main division following formal semantics (intensional vs. extensional), and subclasses of a denotational nature. This resource is the least useful for the purposes of this PhD.

In the MikroKosmos lexicon, one entry per sense is provided. In WordNet, the very notion of *synset* implies that each word potentially participates in several entries. This approach is not feasible in GLiCom's context. Only part of speech and syntactic disambiguation modules are available at GLiCom, so that multiplying entries for semantic reasons is not advisable in the current state of the tools. I will rather use underspecification in case of polysemy (see Section 3.6.4).

Finally, all three resources have been manually developed. Although Nirenburg and Raskin aim at automation, the automation they perform is the propagation of existing entries or parts of entries to related portions of the lexicon, using particular entries as templates (see Nirenburg and Raskin (2004, ch. 9)). For instance, the entries of adjectives *big, small, minuscule, enormous, gigantic* are all the same except for the *range* value of the SIZE attribute. These entries can be automatically created from one seminal entry, and their values manually specified. This kind of approach resembles lexicographic methodology. I will pursue a fully automatic classification, using morphological and syntactic properties as cues to establishing broad semantic classes.

3.6 Parameters to define a semantic classification

We have seen in this chapter very different approaches to the semantics of adjectives, with different goals and methodologies. To see how they fit into the present research, we have to take into account its purpose and motivation. The semantic classification I aim at is subject to constraints of several kinds, listed at the beginning of this chapter.

In what follows, we will review some syntactic arguments that have been used for semantic classification in the literature (Section 3.6.1) and the exploitation of the relationship between morphology and semantics (Section 3.6.2).

3.6.1 Syntactic arguments for semantic classification

The relationship between syntax and semantics for adjectives has been noted in diverse linguistic scholarships. Some of the parameters involved in the syntactic characterisation of adjectives have already been mentioned in the discussion up to now. In this Section, four of them are gathered and related in a more systematic manner to the intended classification: position with respect to the head noun (Section 3.6.1.1), predicative use (Section 3.6.1.2), ordering (Section 3.6.1.3), and coordination (Section 3.6.1.4)

3.6.1.1 Position with respect to the head noun

In descriptive grammar (Section 3.2), it is noted that the position of the adjective with respect to the head noun has consequences in the semantic interpretation: in Catalan, postnominal adjectives trigger a restrictive interpretation (example 3.28), while prenominal adjectives usually trigger a nonrestrictive interpretation (examples in 3.29; all examples adapted from the CTILC corpus).

(3.28) Rebutjava les avingudes **amples** ... per evitar l' esguard dels curiosos
Avoided the avenues wide ... to avoid the look of-the curious

'I avoided wide avenues ... so that curious people would not look at me'

(3.29) a. això que designem amb el **bonic** nom de *neocapitalisme*
that which designate with the nice name of *neocapitalism*

‘that which we give the nice name of “neocapitalism” to’

b. un **dubtós** Che amb una indefinició afectiva cap a Evita
a doubtful Che with an indefinition affective toward to Evita

‘a doubtful Che [Guevara] with an affective indefinition toward Evita.’

c. La **fidel** Armanda viurà a la torre fins que es mori
The faithful Armanda live in the tower until that REFL dies’

‘Faithful Armanda will live in the tower until she dies’

In 3.28, the speaker only avoided avenues that were wide, not those that were narrow. *Amples*, thus, restricts the denotation of the noun *avingudes* to a subset of them that are subject to the constraint specified by the adjective. In 3.29a, the denotation of the noun phrase *el bonic nom de “neocapitalisme”* is the same as *el nom de “neocapitalisme”*, where adjective *bonic* has been omitted. *Bonic* does not restrict the denotation of the noun *nom* to those names that are nice. That is why nonrestrictive adjectives can be used with proper nouns, where the reference is inherently given, as in examples (3.29b-3.29c).

It has also been widely noted that relational adjectives cannot be used in prenominal position in Spanish and Catalan (among others Bosque and Picallo (1996), Demonte (1999), Picallo (2002), McNally and Boleda (2004)). This is probably due to the fact that they are only used in restrictive contexts.

Most intensional adjectives, such as *presumpte* ‘alleged’, can only be used prenominally in Catalan. When an adjective has an intensional reading in addition to an intersective reading, this reading can only be activated in prenominal position. For instance, adjective *antic* means ‘ancient’ or ‘former’ depending on the context. In example 3.30b we see that both orders, pre- and post-nominal, are admitted for the adjective when modifying *manuscrit*. The first is a restrictive reading, the second a nonrestrictive reading, as has been discussed above. When modifying *president*, because the notion of ‘ancient president’ does not make sense, only the prenominal position is admitted, with an intensional reading.

(3.30) a. un manuscrit antic / un antic manuscrit
‘an ancient manuscript’

b. l’antic president / #el president antic
‘a former president’

This restriction does not hold for modal adjectives, such as *possible*, as seen in example 3.31.

(3.31) la dona estava gairebé sempre a casa i tenia com a única professió possible
the woman was almost always at home and had as to only profession possible
la d’atendre les labors domèstiques
the of-attending the labours domestic

‘Women remained almost always at home and had as their only possible job that of attending domestic work’

To sum up, if we use the position of the adjective with respect to the head noun as a criterion for classification, we find support for the intensional (only prenominal), relational (only post-nominal) classes, as opposed to the default (pre- and post-nominal) behaviour.

3.6.1.2 Predicative use

As has been mentioned in Section 3.1, adjectives as a class can be used as predicates in copular sentences and other constructions. At least two kinds of adjectives either cannot or can only in very restricted circumstances: intensional and relational adjectives. The fact that intensional adjectives cannot be used as predicates has been widely noted in formal semantics since the work of Montague. Note, however, that not all semantically intensional adjectives are subject to this restriction: as has been discussed in Section 3.3.2.2, privative and modal adjectives can act as predicates.

Relational adjectives have been deemed nonpredicative (Bally, 1944), or predicative only when used with a different, qualitative reading (Raskin and Nirenburg, 1998). Some authors note that predicative readings are possible with a truly relational reading, but do not specify in what contexts (Demonte, 1999; Picallo, 2002). As has been explained in Section 3.3.2.2, McNally and Boleda (2004) establish that relational adjectives can be used as predicates when their argument is a kind-denoting noun phrase.

Predicative use also singles out intensional and relational adjectives with respect to the default, predicative behaviour typical of adjectives.

3.6.1.3 Ordering

Adjective ordering has been much studied for English, mainly for didactic purposes. The hierarchies that have been built to describe their relative ordering within a noun phrase are denotational in nature. As an example, consider the hierarchy proposed by Picallo (2002) to explain the order of qualitative adjectives, presented in Section 3.2.1.

(Bally, 1944) already noted that relational adjectives are subject to a kind of adjacency constraint: they appear close to the head noun, closer than other kind of adjective or modifer, as exemplified in (3.32) (McNally and Boleda, 2004, ex. (34), p. 189) This is a further syntactic argument for their distinction as a class.

(3.32) a. inflamació pulmonar greu

b. #inflamació greu pulmonar
'serious pulmonary inflamation'

Adjective clustering in prenominal position is much less frequent than in postnominal position. When two or more adjectives occur in prenominal position, intensional adjectives may appear in any order with respect to intersective adjectives, as in example 3.33 (McNally and Boleda, 2004, ex. (22), p. 186). The order, however, affects interpretation ((3.33a) entails that the referent of the noun phrase is young, while (3.33b) does not).

(3.33) a. jove presumpte assassí
'young alleged murderer'

b. presumpte jove assassí
'alleged young murderer'

3.6.1.4 Coordination

Adjectives of the same semantic class tend to coordinate in Catalan, while those of different semantic class are juxtaposed. English is much more prone to juxtaposition even within the same semantic class, which explains the attention devoted to adjective ordering within this language. Example (3.34) is taken from BancTrad, a corpus of translations Badia et al. (2002). The original is in English (3.34a), the translation into Catalan (3.34b) was performed by a human translator independently of corpus collection or research purposes.

(3.34) a. special concrete measures

b. mesures especials i concretes
measures special and concrete

The coordination criterion again supports the delimitation of intensional, intersective and relational adjectives. When modifying the same head, intensional and intersective adjectives do not coordinate, but juxtapose (example 3.35). The same applies to relational and intersective adjectives (example 3.36). This characteristic argues for a difference in semantic function between these classes of adjectives.

(3.35) a. jove presumpte assassí
'young alleged murderer'

b. #jove i presumpte assassí

(3.36) a. inflamació pulmonar greu
'serious pulmonary inflamation'

b. #inflamació pulmonar i greu

Again, modal adjectives are an exception to this rule. They are found in coordination with other modal adjectives (examples in (3.37)), but also with nonintensional adjectives (examples in (3.38)), especially with deverbal *-ble* adjectives that connote potentiality, as *possible* does (see (3.38c)).

(3.37) a. limitar-se a actuar sobre allò més **necessari i possible**
limit-REFL to act upon that most necessary and possible.

'just act upon that which is most necessary and possible.'

b. tocant a la **possible o probable** defensa de Barcelona
concerning to the possible or probable defence of Barcelona

'concerning the possible or probable defence of Barcelona'

c. estalviar trencacolls a algun **possible i futur** historiador.
spare wrecknecks to some possible and future historician

'spare trouble to some possible and future historician'

(3.38) a. ha sotmès els subjectes a un aprenentatge lingüístic, **difícil però possible**
has submitted the subjects to a learning linguistic, difficult but possible

‘he has submitted the subjects to a difficult but possible linguistic learning process’

b. no semblava **prudent ni possible** intentar continuar governant
not seemed advisable nor possible try continue governing

‘It did not seem neither advisable nor possible to try and continue governing’

c. és **possible i desitjable** que els nois més grans siguin conscients de . . .
is possible and desirable that the boys more old be aware of . . .

‘It is possible and desirable for older boys to be aware of . . . ’

3.6.2 Morphology and semantics

As has been noted in the Introduction, the morphology-semantics has received less attention in linguistic theory and NLP than the syntax-semantics interface. However, the relationship between morphological processes and regular meaning effects has been repeatedly noted since the beginning of studies on language. Some morphological processes, such as derivation, have a clear and relatively stable semantic effect.

Derivation serves as the basis for many of the semantic distinctions established in the literature for adjectives (thus exploiting the morphology-semantic interface), as reported in this chapter. The only tradition that has ignored this level of analysis in classifying adjectives is formal semantics. A summary of the classifications and their relationship to morphology (if specified) follows.

descriptive grammar: qualitative (not specified) / relational (denominal adjectives) / adverbial (adverbialising adjectives)

formal semantics: intersective / subsective / intensional

ontological semantics: scalar (not specified) / object-related (denominal adjectives) / event-related (deverbal adjectives + “suppletivists”)

NLP: WordNet: descriptive (not specified) with subclass participial (deverbal) / relational (denominal adjectives)

SIMPLE: similar to formal semantics

MikroKosmos: the same as ontological semantics

Most proposals highlight denominal adjectivisation as originating a distinct type of adjectives (relational or object-related adjectives). Two proposals (ontological semantics and WordNet) signal deverbal adjectivisation as a similar process, namely, productively creating a class of adjectival meanings from the meanings of verbs. Only in descriptive grammar a class is suggested that results from the property of being able to produce other parts of speech: adverbial adjectives are typically those to which an adverbialising *-ment* suffix can be attached to produce an adverb (*freqüentment* ‘frequently’, *ràpidament* ‘quickly’; see Section 3.2.3).

Note, however, that because of the mismatches that arise in the morphology-semantics mapping, no isomorphism can be assumed. Mainly due to diachronic changes, morphological processes almost always have an "irregular" part to them. Once a word is created with a morphological process, it becomes a word of the language, and can undergo further semantic changes. It is the phenomenon known as *lexicalisation*.

An example is useful here. As has been noted in Section 3.4 above, deverbal adjectives ending in *-ble* can usually be paraphrased as "that can be V-ed": for instance, *replaceable* can be paraphrased as "that can be replaced". Some adjectives that were once subject to this productive process have acquired a life of their own. In Catalan, which has the same *-ble* suffix as English, one such case is *amable*. The verb *amar*, 'love', from which *amable* derives is no longer in use in standard modern Catalan (it has been replaced by *estimar*). The adjective no longer means "that can be loved", but 'friendly, nice'. It is, thus, a qualitative or intersective kind of meaning.

The reverse case also exists. These are cases where the root or the suffix are not used in the language, but the word still has a "derived" meaning. Cultisms are an abundant source of this kind of mismatch between morphology and semantics. Latin or Greek words were introduced into Catalan (and other languages) in the XVIII and XIXth centuries, mainly for the rapidly growing scientific terminology. For instance, adjective *bèl·lic* from Latin *bellicus* ('of war, warlike').

Syntax could provide a better clue for these cases than morphology. For instance, if an adjective has acquired a meaning different to the "compositional" meaning that would result of the productive application of a morphological process, it will not behave like adjectives that have the derived reading. This hypothesis will be tested in Chapters 5 and 6.

3.6.3 Classification

One of the purposes of this PhD is to define a classification for adjectives that is subject to the constraints explained in this chapter and at the same time takes into account the insights provided in the literature on adjective semantics in diverse traditions, as well as implementations in NLP resources. However, another main goal is to establish inductive mechanisms in the definition process, so that hypotheses are tested and can provide feedback for the classification.

As a result, the targeted classification has changed during the process of the PhD. The first version of the classification (Boleda and Alonso, 2003; Boleda, 2003) distinguished between intensional, relational, and qualitative adjectives (this class covering subsective and intersective adjectives as termed in formal semantics). Semantic and syntactic arguments were used for this classification, following the lines of Sections 3.2, 3.3 and 3.6.1. The classification is a blend of insights from descriptive grammar and formal semantics.

As will be seen in Chapter 5, a series of unsupervised acquisition experiments, together with some theoretical considerations, led to the abandoning of the intensional class and the introduction of a third, event-related class. This class receives theoretical support from ontological semantics (Section 3.4). The second version of the classification responds better to the desiderata explained in this Section in the sense that it is based on a single parameter (the ontological type of the denotation) and it is more balanced (classes are of a similar size). It will raise new problems and research questions, as will be discussed throughout this document.

Table 3.4 summarises the two classifications proposed. Note that the terminology changes from version I to version II of the classification, for consistency reasons.

Classification A	**Classification B**	**Examples** ***Translation***
qualitative	basic	vermell, rodó, autònom, subaltern *red, round, autonomous, auxiliary*
relational	object-related	pulmonar, estacional, botànic *pulmonar, seasonal, botanical*
intensional	–	presumpte, antic *alleged, former* (in one sense; *ancient* in another)
–	event-related	abundant, promès, variable *abundant, promised/engaged, variable*

Table 3.4: *Two proposals for adjective classification.*

The characteristics of each of the two classifications will be reviewed in the relevant Sections of chapters 5 and 6.

3.6.4 Polysemy

As any semantic classification, our classification is affected by polysemy. However, the kind of polysemy involved is just a subset of the kinds of polysemy examined in the literature, namely, polysemy affecting class distinctions. We will only consider sense distinctions that involve different classes, typically between a qualitative or basic reading and the rest, as the following examples illustrate:

(3.39) a. edifici antic / antic president
building ancient / former president
'ancient building / former president'

b. reunió familiar / cara familiar
meeting familiar / face familiar
'family meeting / familiar face'

c. conseqüència sabuda / home sabut
consequence known / man wise
'known consequence / wise man'

Antic (example (3.39a)) has two major senses, one qualitative or basic (equivalent to 'old, ancient') and the other intensional (equivalent to 'former'). Note that when used in the intensional sense, it appears prenominally, as discussed in Section 3.6.1.1 above. *Familiar* (example (3.39b)) also has two major senses, one relational or object-related (which in English is typically translated with an attributive noun, *family*), and one qualitative or basic (equivalent to 'familiar'). Similarly, *sabut*, participle of the verb *saber* ('know') has the expected resultative sense but also a qualitative or basic sense, equivalent to 'wise'.

In all these examples, the qualitative sense participates of all the syntactic environments typical of the class (discussed in Section 3.6.1), and are gradable, as is shown in examples (3.40-3.42).

(3.40) a. edifici molt antic / #molt antic president
building very ancient / very former president

'very ancient building / #very former president'

b. Aquest edifici / #president és antic
This building / #president is ancient

'This building / #president is ancient'

(3.41) a. #reunió molt familiar / cara molt familiar
#very meeting familiar / face very familiar

'#very/much family meeting / very familiar face'

b. Aquesta cara / #reunió em resulta familiar
This face / meeting to-me results familiar

'This face / #meeting is familiar to me'

c. Em va rebre la familiar cara de la Maria
Me PAUX-3ps welcome the familiar face of the Maria

'The familiar face of Maria welcomed me'

d. #La familiar reunió era molt animada
The familiar meeting was very lively

'The family meeting was very lively'

(3.42) a. Aquest home és molt sabut!
This man is very wise

'This man is very wise!'

b. #Aquesta conseqüència és molt sabuda!
This consequence is very known

'This consequence is very known'

The basic readings, but not the intensional, object- or event-related readings, yields in general gradable adjectives (examples (3.40a), (3.41a), (3.42a)). Similarly, it allows predicative constructions, in copular sentences (examples (3.40b), (3.42a)) or other predicative environments (example (3.41b)). It also allows pre-nominal position of the adjective, which is not possible for the object reading (example (3.41c)).

Other kinds of polysemy that have traditionally been tackled in the literature will not be considered, as they do not affect the class distinctions drawn in Section 3.6.3. For instance, we will not be concerned with the following kinds of polysemy.

(3.43) a. discurs llarg / carrer llarg
speech long / street long

'long speech / long street'

b. noi trist / pel·lícula trista
boy sad / film sad

'sad boy / sad film'

c. aigua clara / explicació clara
water clear / explanation clear

‘clear water / clear explanation’

d. visita papal / cotxe papal
visit Papal / car Papal

‘Papal visit / Papal car’

e. conducta abusiva / dictador abusiu
behaviour abusive / dictator abusive

‘abusive behaviour / abusive dictator’

Examples (3.43a-3.43b) involve the kind of parameter modified (time or physical property in case of *long*; individual argument or event argument in the case of *sad*; see Pustejovsky (1995, p. 127 ff.)). The polysemy in (3.43c) is a typical metonymy case in which physical properties are used in an abstract domain. In examples (3.43a-3.43c), the class of the adjective is qualitative in both senses.

As for examples (3.43d) and (3.43e), they involve the semantic role of the noun. Example (3.43d) has been discussed in Section 3.2.2 (example 3.12). As for (3.43e), it is argued in Raskin and Nirenburg (1996, p. 108) that “What is abusive is either the event (E) itself, as in *abusive speech* or *abusive behaviour,* or the agent (A) of the event, as in *abusive man* or *abusive neighbor*”. It is thus an analysis along the lines of Larson’s (1998) analysis of *beautiful* and Bouillon’s (1999) analysis of *vieux* explained in Section 3.3.2. In both readings, the adjective is object- (example (3.43d)) or event-related (example (3.43e)).

Note that all these alledgedly adjectival cases of polysemy are similar in that the polysemy effect has more to do with the semantics of the modified noun than that of the adjective.

In some cases, the sense distinctions are not as clear-cut as the cases examined in 3.39. For instance, Raskin and Nirenburg (1998) note that it is not easy to draw a line between gradable and non-gradable relational adjectives, and we have seen in (3.41) above that gradable uses of relational adjectives are typical of a shift to a basic reading. They show that it is very hard, if not impossible, to find a relational adjective that cannot be used in a qualitative sense. For instance, although “it is hard to imagine a more truly-[relational] adjective than *aeronautical* ‘related to aeronautics’ ”, examples like those in (3.44) can be constructed (Raskin and Nirenburg, 1998, ex. (42), p. 173).

(3.44) a. His approach to the problem was aeronautical.

b. His approach to the problem was much more aeronautical than mine.

As the authors argue, it seems that a productive semantic process takes place along the lines of schema (3.45) (Raskin and Nirenburg, 1998, schema (43), p. 173) so that all relational adjectives can be used as scalars. The degree of lexicalisation and actual use of this possibility is different for different relational adjectives, and eventually yields totally differentiated senses, such as ‘known, familiar’ for *familiar* or ‘cheap’ for *econòmic*.

(3.45) Pertaining to [noun meaning] → characteristic of [noun meaning]

A paradigmatic case are relational adjectives that can be predicated of humans, such as nationality-denoting adjectives or “social property” (as termed in SIMPLE; 3.5.2) adjectives. They can be viewed as properties when modifying human-denoting heads, and indeed in these cases they can be predicatively used (example (3.46)). However, when not modifying human referents, they are closer to true relational adjectives. They do not follow the usual inferential patterns for intersective adjectives: (3.47a) does not entail (3.47b). In fact, the predicative construction in (3.47b) is anomalous.

(3.46) El Paul és alemany/catòlic/comunista
The Paul is German/Catholic/communist

‘Paul is German/Catholic/ a communist’

(3.47) a. El Reichstag és el parlament alemany
The Reichstag is the parliament German

‘The Reichstag is the German parliament’

b. ?#El Reichstag és alemany
The Reichstag is German

Some authors ((Peters and Peters, 2000), (Carvalho and Ranchhod, 2003), (Bohnet et al., 2002)) posit an additional class for some or all of these cases. We will treat them as a special case of the object-related vs. basic polysemy, although they are probably best viewed as underspecified, rather than polysemous.

3.7 Adjectives and Lexical Acquisition

We focus in this section on Lexical Acquisition research on adjectives. For overviews of research on Lexical Acquisition for verbs, see, e.g., McCarthy (2001); Korhonen (2002b); Schulte im Walde (2003).

Adjectives have received much less attention than nouns and especially verbs in Lexical Acquisition. Hatzivassiloglou and McKeown (1993) was one of the first pieces of research oriented toward acquiring sets of semantically related adjectives. Their purpose, however, was not automatic classification, but identification of adjectival scales from corpora. They used adjective-noun and adjective-adjective cooccurence frequency to determine similarity of meaning for 21 lemmata,and they clustered them. Adjective-noun occurences were considered to be positive information, following the hypothesis that two adjectives that consistently modify the same nouns have related meanings, and may belong to the same scale. Adjective-adjective occurences were used as negative information, as usually two adjectives that are concatenated have different meanings and hence do not belong to the same scale.

Information on gradability or the qualitative/relational distinction was not taken into account; the authors themselves stated the need to include these pieces of information. In a later paper, Hatzivassiloglou and McKeown (2000) present a statistical model to classify adjectives according to gradability that obtains a very high precision (87.97%) using a very simple indicator, namely occurence after a degree modifier.

Hatzivassiloglou and McKeown (1997) seeked to automatically identify semantic orientation, that is, within scalar adjectives, which are oriented toward the positive pole and which toward the negative pole of the scale. They used coordination information. The log-linear regression model they used predicted whether two coordinated adjectives were of the same or different orientation. A clustering procedure determined the positive or negative orientation of every set of adjectives.

The same kind of information (coordination) was used in Bohnet et al. (2002) for different, more traditional classification purposes. They aimed at a classification defined in a German descriptive grammar (Engel, 1988) between quantitative ones (similar to determiners, like *viele, einige* ‘many, some’), referential ones (*heutige*, ‘of today’), qualitative ones (equivalent to our own qualitative adjectives), classificatory ones (equivalent to relational adjectives), and adjectives of origin (*Stuttgarter*, ‘from Stuttgart’).

They applied a bootstrapping approach. The procedure is as follows: start from a manually classified set of adjectives, detect coordinations with an already classified member, assign the same class to the other member of the coordination, and iterate the procedure with the newly annotated adjectives, until the algorithm does not produce any more changes in the classification. This procedure has two basic problems: first, that it assigns a single class to each adjective, and so it is not possible to identify adjectives belonging to more than one category. Second, that coordination data only apply to a reduced number of lemmata, because most lemmata that occur in a given corpus do not occur in coordinating constructions. To alleviate these two problems, the authors applied another algorithm that exploits the order of adjective classes within an NP (in German, quantitative $<$ referential $<$ qualitative $<$ classifying $<$ of origin). This algorithm and its results, however, is not explained in detail.

A kind of classification that bears similarities with our own and with the classification in Bohnet et al. (2002) was used in Carvalho and Ranchhod (2003) to disambiguate adjective and noun readings in Portuguese. The classification included the following classes and subclasses: predicative color adjective, nationality denoting adjective, predicative adjective, predicative adjective (specifying whether only postnominal or both pre- and postnominal positions are possible when acting as modifiers), relation adjectives, and adjectives with determinative value. [28] Carvalho and Ranchhod manually coded 3,500 adjectives with this information, and built a series of finite state transducers to model noun phrases and disambiguate between nominal and adjectival readings. Adjective information served to establish constraints within the transducers, for instance that adjectives of different classes cannot coordinate.

In a recent paper, Yallop et al. (2005) have aimed at the acquisition of syntactic subcategorisation patterns for English adjectives. They identify over 30 adjectival subcategorisation frames, including detailed information such as the nature of the arguments of predicative adjectives (finite and non-finite clauses, noun phrases). They use a statistical parser to extract grammatical relations and pattern matching to hierarchically classify these grammatical relations into frames.

Beside these pieces of research with a classificatory flavour, there have been other lines that exploited lexical relations among adjectives, mostly polysemy and antonymy, for Word Sense Disambiguation. Justeson and Katz (1995) consider polysemous adjectives with more than one antonym, for instance *old*, with antonyms *young* and *new*. They use the nouns that adjectives modify as clues for the disambiguation of the adjective sense. To determine which nouns corre-

[28] The 3 classes of predicative adjectives are lumped together in our classifications as qualitative or basic. Nationality denoting and relation adjectives are our object-related adjectives. Adjectives with determinative value have been termed intensional adjectives here. Note that the authors claim their classification to be syntactic, but it is clear from the labels that it has some semantic content. Like our classification, it is at the syntax-semantics interface.

spond to which sense, they examine the antonyms for the relevant senses. For instance, *man* is modified by *old* and *young*, and *house* by *old* and *new*. When the occurence *old house* is found, it can be determined that *old* is used in the 'not new' sense, not in the 'aged' sense. This work was based only in 5 adjectives (*hard, light, old, right, short*) and the corpus used for acquisition was manually disambiguated.

Chao and Dyer (2000) tackle the same task in a more sophisticated fashion: they analyse 135 adjectives and use sources of information with little manual annotation. They build a Bayesian classifier that uses information from WordNet (as source for semantic information) and the Web (to tackle data sparseness issues) so as to disambiguate adjectives such as *great*: note that when occuring in *great hurricane* it means 'strong', not 'good' as in many other contexts.

Lapata (2000; 2001), as opposed to studies mentioned so far, focuses on the meaning of adjective-noun combinations, not on that of adjectives alone. In the Generative Lexicon framework (Pustejovsky, 1995; Bouillon, 1997), she attempts at establishing the possible meanings of adjective-noun combinations, and at ranking them using information gathered from the British National Corpus (Burnage and Dunlop, 1992). This information should indicate that an *easy problem* is usually equivalent to *problem that is easy to* **solve** (as opposed to, e.g., *easy text*, that is usually equivalent to *text that is easy to read*). She explored nine adjectives, and ten noun-adjective combinations for each of the adjectives.

Lapata (2000) extracts noun-adjective pairs and identifies noun-verb and verb-adjective/adverb pairs that are related with every noun-adjective pair. The relevant noun-verb pairs are those in which the noun is the same as in a particular noun-adjective pair. The relevant verb-adjective/adverb pairs are those in which the adjective or adverb modifying the verb is the same as the adjective under investigation, or is morphologically derived from it. For instance, for *easy problem* verbs are identified which occur with *problem* as subject or object, as well as verbs modified by *easy* or *easily*. The informations are crossed so as to obtain the following verb ranking for *easy problem*: *solve, deal with, identify, tackle, handle* (Lapata, 2000, p. 151).

The author, thus, builds a probabilistic model from the corpus and computes the probability of a paraphrase from the cooccurence frequencies of the relevant noun-verb and verb-adverb pairs. Note that in this case there is no pre-defined sense inventory, contrary to research in Justeson and Katz (1995; Chao and Dyer (2000).

Most of the work done on Lexical Acquisition for adjectives (in fact, all except for Bohnet et al. (2002)) focuses on different phenomena and tasks than that tackled here, because they neither establish an semantic classification of adjectives nor attempt at acquiring class polysemy. The task is different: they either try to infer aspects of the organisation of adjectives around scales (research by Hatzivassiloglou and colleagues), which affects a small subset of the adjectives, argued in Section 3.5.1, or are oriented toward disambiguation (either Word Sense Disambiguation, as in Justeson and Katz (1995); Chao and Dyer (2000) , or part of speech disambiguation, as in Carvalho and Ranchhod (2003)), or infer paraphrases for particular noun-adjective combinations (Lapata, 2000). Yallop et al. (2005) tackle syntactic, not semantic classification.

Some of the pieces of research reviewed in this Section deal with polysemy. However, they pursue a different kind of polysemy than the polysemy aimed at here. Work in Word Sense Disambiguation typically has to do with polysemy associated with selectional restrictions, like the *old-young / old-new* dychotomy mentioned above. Work in the Generative Lexicon framework analyses the kind of adjectival polysemy that is related to the structure of the modified noun, following the Generative Lexicon theory (Lapata, 2000), similarly to the *beautiful dancer* example discussed in Section 3.3.2.1. However, many methodological aspects and insights are related and are used in several ways for the purposes of the research presented here, as will become clear in the following chapters.

3.8 Summary

This chapter has provided a general review of adjectives as a part of speech and the treatment of their semantics in several linguistic traditions, as well as a summary of previous related work in Lexical Acquisition. The classification proposed in this PhD is the result of this review, together with empirical data gathered from machine learning experiments (see Chapter 5).

Formal semantics may seem the natural tradition to look for a semantic classification of adjectives, because in this field some requirements for Computational Linguistics, such as an explicit formalisation of semantic analyses, are met. However, we have argued that formal semantics does not provide an adequate account of lexical semantics in general, and adjective semantics in particular, at least for the purposes of this PhD. Moreover, much research in formal semantics has focused on a relatively minor class, intensional adjectives, which causes coverage problems for NLP systems. Also, the parameters explored in this tradition (mainly entailment patterns) are very difficult to use as criteria for classification. Finally, the classes that emerge from a straightforward application of the meaning postulates discussed in Section 3.3.2 are formed by heterogeneous subclasses, which questions the usefulness of the distinctions.

Standard descriptive grammars such as Picallo (2002) for Catalan, as well as proposals used in NLP resources, do propose broad-coverage classifications. In fact, one of the main distinctions made in these proposals (qualitative vs. relational) is central to this PhD.

However, the definition and characterisation of the classification, in its final version, corresponds most closely to the one formulated within the Ontological Semantics framework, most notably in Raskin and Nirenburg (1998). Within the PhD process, we formulated a first attempt at a classification proposal that was a hybrid between formal semantics and descriptive grammar, distinguishing between intensional, qualitative and relational adjectives. The parameters used in this classification were not at the same level: intensional adjectives are defined in terms of their entailment behaviour, while relational adjectives are defined through the ontological kind of their denotation (qualitative adjectives lie somewhat in between).

The second version of the classification establishes a unique classification parameter, namely, the ontological sort of adjectival denotation, or the kinds of meaning adjectives can have. While all adjectives can be said to denote properties, these properties can be instantiated as simple attributes (basic adjectives), relationships to objects (object-related adjectives), or relationships to events (event-related adjectives). This classification meets most of the constraints established in Section 3.6: it is broad in coverage, balanced, and consistent. It draws on the morphology-semantics and syntax-semantics interface, while still being defined in purely semantic terms. Because of its use of these interfaces, it is amenable to automatic acquisition, as will be shown in the following chapters. Some of the phenomena discussed in this chapter (modality, some distinctions in the semantic sort of adjective arguments) are not covered with this classification, and are a subject for future research.

The classification presupposes the definition of an ontology as a model of the world, anchoring meaning in external reality. Despite the numerous philosophical and practical problems connected to defining such an ontology, we believe that only an approach that explicitly models reality (however this model is instantiated) has a chance of achieving explicative and predictive power in analysing the semantics of natural languages. This is even clearer when dealing with Computational Linguistics: if computers are to simulate understanding and engage in productive linguistic interactions, the concepts that humans share and their interrelationships have to be modeled in a formal language.

The semantic analysis of relational adjectives summarised in Section 3.3.2.1 has been published in the following article:

McNally, L. and Boleda, G. (2004). Relational adjectives as properties of kinds. In Bonami, O. and Hofherr, P. C. (Eds.), *Empirical Issues in Syntax and Semantics 5*, pages 179–196. http://www.cssp.cnrs.fr/eiss5/.

Chapter 4
Gold Standard

> [Web experiments hold] the promise to achieve . . . methodological and procedural advantages for the experimental method and a previously unseen ease of data collection for scientists and students.
>
> Reips (2002, p. 243)

This chapter explains various efforts devoted to achieving a reliable Gold Standard for the classification proposed in Chapter 3, for use in the machine learning experiments explained in Chapters 5 and 6. No such Gold Standard is available in the literature, because in theoretical work only a few examples for each class are mentioned, and the classifications discussed in different works do not exactly match our classification proposal. Therefore, it is necessary to gather human judgments with respect to the semantic classes that we want to obtain.

If there is low human agreement on the class of each adjective, the Gold Standard has low reliability: it can not be used as a basis for any kind of decision, including machine learning experiments. The relationship between agreement and reliability is summarised in Krippendorff (2004b, p. 414) as follows: "agreement is what we measure; reliability is what we wish to infer from it." Although agreement in all Gold Standard experiments we report is far greater than chance, we have not succeeded in achieving high agreement, most notably for polysemy judgments. A reliable Gold Standard is very difficult to achieve in areas such as lexical semantics (Merlo and Stevenson, 2001).

4.1 Initial efforts

The establishment of a Gold Standard has evolved during the PhD, in parallel to the evolution of the classification. The first classification proposal distinguished between qualitative, relational, and intensional adjectives, and additionally, two polysemous classes (polysemous between qualitative and relational, or qualitative and intensional). In the first experiments, explained in Section 5.1, we used a 101 unit Gold Standard, randomly chosen among all adjectives occurring more than 10 times in the corpus. 50 lemmata were chosen token-wise from the corpus and 50 type-wise from the lemma list (a lemma was chosen with the two methods, and one repetition was removed). Two more lemmata were added because the intensional class was otherwise not represented (recall from Section 3.3 that intensional adjectives are a very small class).

The lemmata were annotated by 4 PhD students in Computational Linguistics. The task of the judges was to assign each lemma to one of the five classes mentioned in the previous paragraph. The instructions for the judges included information about all linguistic characteristics discussed in Chapter 3, including entailment patterns, other semantic properties, and morphosyntactic characteristics. The judges had a moderate level of agreement (kappa 0.54 to 0.64; see discussion on agreement coefficients in Section 4.3). They reported difficulties using entailment patterns as classification criteria, as well as tagging particular kinds of adjectives, such as deverbal adjectives. The fact that the judges received information concerning expected syntactic behaviour is not optimal: if the idea is to automatically obtain semantic classes on the basis of morphosyntactic information, the human classification should be purely semantic. This aspect was corrected in further Gold Standard experiments.

As a result of the unsupervised experiments explained in Section 5.1, the classification was refined: the final classification proposal distinguishes between basic, event-related, and object-related adjectives, as mentioned in Section 3.6. For subsequent experiments (Section 5.2), we developed an 80 unit Gold Standard, randomly chosen in a token-wise fashion, this time establishing a higher frequency threshold (50 times).

The 80 lemmata were independently annotated by three PhD students in Computational Linguistics (two of which, including the author of this thesis, carry out research on adjectives). The task was to classify each adjective as either basic, event-related or object-related. [1] The judges received instructions which referred only to semantic characteristics, not to the expected syntactic behaviour. For example, so as to detect event-related adjectives, "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 classified.

The judges were allowed to assign a lemma to a second category in case of polysemy. They were asked to assign the first class to the most frequent sense and the second class to the least frequent sense, according to their intuitions. In these experiments, thus, polysemous items were not assigned to separate classes as in the first Gold Standard experiment.

The agreement scores for the main class were quite high for a lexical semantic task: kappa 0.68 to 0.80. However, the agreement scores for polysemy judgments were not significant at all: the judges did not even agree on whether an adjective is polysemous or not, let alone the particular classes. Polysemy judgments were not used in the classification experiment because of their unreliability. The three main class classifications were merged into a single Gold Standard set for the machine learning experiments (see Chapter 5).

From these two experiences, a number of conclusions were drawn, concerning the frequency threshold, the number of lemmata and judges needed to reach stable conclusions, the sampling methodology, the design of the classification task for human judges, and the analysis of results. These shaped the final Gold Standard experiment presented in the remaining of this chapter.

The experiment was aimed at eliciting semantic classes using judgments from native speakers, and its goals were the following:

- to build a robust Gold Standard for use in the ML experiments
- to assess its replicability
- to achieve insight into polysemy
- to detect conflictive lemmata or classes

[1] The judges also classified the lemmata along an additional parameter, arity, or the number of arguments of the adjective. This parameter is not relevant here and will not be further discussed.

- more generally, to provide feedback to the proposed classification of adjectives

To meet some of these goals, 3-4 judges do not provide enough data. An example is the distinction between "difficult" and "easy" adjectives: if a large number of judges is involved in the classification process, different degrees of difficulty can be expected to more or less correspond to the degree of agreement with respect to their class. The source of the difficulty could lie in their semantics not fitting in the classification, or in some other characteristic. Polysemy judgments also proved difficult to analyse with few judges, again because no trends in their distribution could be observed.

For these reasons, we designed an experiment via web. Web experiments are a very powerful tool to gather linguistic data. They have recently started to be used for psychological research, so that even guidelines for conducting such web experiments have appeared in psychology journals (Reips, 2002). In psycholinguistics, a number of studies that gather data through the Web have appeared (Lapata et al. (1999), Corley and Scheepers (2002) , Melinger Schulte im Walde (2005), among others). Keller et al. (1998) have developed software specifically for that type of experiments.

Web experiments allow higher number and variety of data to be gathered than traditional, laboratory based experiments, at virtually no cost. Voluntary participation is much easier to achieve because no trips or appointments are involved. However, they also raise a number of difficulties. One such difficulty is the fact that a web experiment restricts the kind of participant to people with internet connection and some computational skills. Reips (2002) claims that the gain in number and variety of participants overweights this limitation. Moreover, in most cases (as our own case) the computational expertise required is low (basic knowledge of browser use). Another difficulty is that the experimenters lose control over the environment in which the experiment takes place, and participants' anonymycy makes it sometimes difficult to assure the quality of the data. Reips (2002) offers some guidelines to avoid or soften these problems. Despite these difficulties, we decided to carry out a web-based experiment, so that we could easily recruit participants without any restrictions on time or place, and with no economic cost

Because we wanted a large amount of judgments about adjectives, we decided to address the experiment to the general population, not limiting it to expert judges. The use of naive subjects for linguistic tasks is not uncommon in Computational Linguistics research. For instance, Fellbaum et al. (1998) compare naive and lexicographer judges in the task of tagging a text with WordNet senses. Recently, Artstein and Poesio (2005a) use 18 naive subjects for coreference tagging. Our experiment only required participants to be minimally educated with respect to linguistic notions (mainly, the concepts *noun*, *verb*, and *adjective*, and familiarity with dictionaries, as will become clear in Section 4.2). This educational level is usually reached in primary school.

Our goal is not only to evaluate the agreement level reached by humans when performing the task, but also to assess our classification proposal and to gain insight into the semantics of adjectives. Thus, we use the process of establishing a Gold Standard as a further empirical empirical tool to gain insight into the semantic classification of adjectives.

We next address each of the aspects of the experiment: method (Section 4.2), the assessment of interrater agreement (Section 4.3), analysis of results (Section 4.4), and analysis of the sources of disagreement (Section 4.5).

Band	#Lemmata	Range	Mean	Mean/million
Low	921	50-121	79.8	5.5
Medium	767	121-328	195.9	13.5
High	603	328-19,545	1146.5	79.1

Table 4.1: *Frequency bands for the stratified sample.*

4.2 Method

4.2.1 Materials

We selected 210 lemmata from the database developed by Roser Sanromà, considering only lemmata with at least 50 occurrences in the study corpus (see Section 2.1.2 for details about the database). This roughly corresponds to 10% of the database, and is more than double as many as the previously constructed Gold Standards. The selected lemmata should be representative of adjectives in Catalan. We consider three factors of variability we want to account for in our analysis, and have therefore to be considered while building the Gold Standard: frequency, morphological type, and suffix.

As has been discussed in Section 2.1.3, adjective frequency presents a Zipfean distribution (with a low number of highly frequent items and a large number of unfrequent items). So as to achieve a balance between the two extremes, we divided the frequency into three equal bands and randomly selected a pre-specified number of lemmata from each band. Instead of considering raw frequencies, however, we took log-transformed frequencies, so as to smooth the curve. The same procedure was followed by Lapata et al. (1999) to choose material for plausibility ratings concerning adjective-noun combinations.

Following this methodology, the thresholds of the frequency bands were set at 121 and 328 occurences. Further information about the frequency bands is depicted in Table 4.1: number of lemmata in each band, range of their frequencies, mean frequency, and mean frequency per million (dividing mean by 14.5 million words in the corpus). Note that the difference in the number of lemmata per frequency band would be even higher if lower-frequency lemmata were considered, again due to the Zipfean distribution of adjectival frequencies.

As for the morphological factors, the derivational type (whether an adjective is denominal, deverbal, or not derived) is not evenly distributed: as shown in Table 2.5 in page 12, there are 399 deverbal lemmata, as opposed to, e.g., 860 denominal lemmata. Moreover, the distribution of lemmata is particularly skewed with respect to the suffix within each of the denominal and deverbal groups. The distribution is shown in Table 4.2. In this table, the first column shows the number of lemmata bearing each of the suffixes for deverbal (V) adjectives, and the 3 remaining ones for denominal (N) adjectives. Adjectives grouped under the suffix group *other* correspond to infrequent suffixes lumped together in Sanromà (2003).

One of the aims of the research is to explore the relationship between morphology and semantics. In addition, a reasonable hypothesis is that one of the sources of semantic variability for adjectives is precisely morphological variability. In previous work, where raw random selection was performed, we found that there were too few lemmata of some morphological types for some of the suffixes or morphological types, so that no analysis could be carried out with respect to the morphology-semantics mapping for them. Therefore, we also designed a stratified approach to morphology, and took an (approximately) equal number of lemmata from each morphological type and from each suffix. The exception were suffixes with very few lemmata

V		**N**					
t (part.)	519	ic	256	*í	19	*ià	4
nt	140	al	230	*iu	18	*íac	3
ble	109	ós	86	*er	16	*ífic	3
iu	91	ari	57	*i	14	*ut	2
or	35	ar	33	*ès	12	*aci	1
*ori	12	ista	31	*at	11	*esc	1
*ós	7	à	25	*ístic	9	*ívol	1
*er	5	**other*	23	*enc	5		

Table 4.2: *Distribution of lemmata across morphological type and suffix.*

(those marked with an asterisc in Table 4.2), which were lumped together in one group.

The distribution of the data in the final selection of lemmata for the Gold Standard is shown in Table 4.3. As can be gathered from the table, the criterion is to have an equal distribution among morphological types (not derived, denominal, and deverbal; 70 lemmata each) and frequency bands (approximately 70 lemmata for each band). Note that participial adjectives are not considered separately, but as a subset of deverbal adjectives (marked with suffix -*t (part.)*).

Within each morphological type, equal distribution among suffixes has also been attempted at. Minor deviations from the expected values are due to either a particular suffix not having enough lemmata for a particular frequency band (e.g., suffix -*or* for the higher frequency band) or to the need of all values summing up to 70 (e.g., deviations for suffixes -*à* or -*t (part.)*).

Morph. type	**Suffix**	**Low**	**Medium**	**High**	**Subtotal**	**Total**
not derived	-	23	24	23	70	70
denominal	à	3	2	2	7	
(N)	al	3	3	3	9	
	ar	3	3	3	9	
	ari	3	3	3	9	
	ic	3	3	3	9	
	ista	3	3	3	9	
	ós	3	3	3	9	
	other	3	3	3	9	
	total (N)					70
deverbal	ble	4	4	3	11	
(V)	iu	4	4	3	11	
	nt	4	4	3	11	
	or	4	5	2	11	
	t (part.)	5	5	5	15	
	other	4	5	2	11	
	total (V)					70
total		72	74	64	210	210

Table 4.3: *Stratification of the Gold Standard.*

We wanted our experiment to last about 30 minutes on average. 210 lemmata is clearly too high a number of lemmata for that duration. Therefore, our 210 Gold Standard dataset was randomly divided into 7 test sets with 30 lemmata each, and each participant would examine only one test set, in the fashion that will be explained in the next section.

4.2.2 Design of the experiment

Our goal was to classify adjectives as basic, event, or object, taking polysemy into account. However, instead of directly asking participants to classify adjectives, and in order to define a task as intuitive as possible (given the problem), we asked the participants to **define** adjectives according to pre-defined patterns, each corresponding to a semantic class. Each definitional pattern acts as a paraphrase that should apply if adjectives fit in one of the classes foreseen in the classification. We thus gather judgements of native speakers with respect to paraphrase relationships between lexical items. Paraphrases are one of the types of linguistic evidence mostly used by semanticists in their research (Katz, 1972), (Lappin, 1996), (Chierchia and McConnell Ginet, 2000).

The task of the participants was to complete the definition for each adjective by filling in a blank field corresponding to a noun, verb, or adjective, depending on the definitional pattern. Filling in a field (signalled as $\boxed{}$ in what follows) implies selecting a definitional pattern and thus a particular kind of meaning, or semantic class. The fact that participants had to fill in the blank instead of simply selecting the pattern made sure that they would pay attention to the task, and also served analysis purposes, giving a clue as to which sense is being signalled in each case. Each field was accompanied by an indication of the expected part of speech (adjective, noun or verb), so as to further constrain the task.

For basic adjectives, the definitional pattern should be filled with a synonym or an antonym, for many basic adjectives have lexical antonyms or near-antonyms, even if not all respond to this lexical relationship, as has been discussed in Chapter 3. The definitional pattern is reproduced in (4.1a) and exemplified in (4.1b).

(4.1) a. Té un significat semblant a / contrari a $\boxed{}_{(adjectiu)}$
‘Has a meaning similar to / opposite to $\boxed{}_{(adjective)}$’
b. **gran** → Té un significat semblant a / contrari a $\boxed{\text{petit}}_{(adjectiu)}$
‘big → Has a meaning similar to / opposite to $\boxed{\text{small}}_{(adjective)}$’

For object-related adjectives, the definitional pattern or paraphrase expressed the relationship to an object lexicalised through a noun, thus reproducing the generic PERTAIN-TO schema of Ontological Semantics represented in Figure 3.3, page 47, as shown in (4.2).

(4.2) a. Relatiu a o relacionat amb (/el/la/els/les/l’) $\boxed{}_{(noun)}$
‘Related to (the) $\boxed{}_{(noun)}$’
b. **bèl·lic** → Relatiu a o relacionat amb (/el/la/els/les/l’) $\boxed{\text{guerra}}_{(noun)}$
‘bellic → Related to (the) $\boxed{\text{war}}_{(noun)}$’

For event-related adjectives, the definitional pattern expressed the relationship to an event lexicalised through a verb. Three definitional patterns were provided to account for the different meanings arising from different suffixation processes: an “active” meaning for suffixes such as *-iu* or *-or* (pattern in (4.3)), a “passive” meaning for the *-ble* suffix (pattern in (4.4)), and a resultative meaning for participial adjectives (pattern in (4.5)).

(4.3) a. que $\boxed{}_{(verb)}$
‘that/which/who $\boxed{}_{(verb)}$’

b. **constitutiu** → que $\boxed{\text{constitueix}}_{(verb)}$
'constitutive → that/which $\boxed{\text{constitutes}}_{(verb)}$'

(4.4) a. que pot ser □$_{(verb)}$
'that can be □$_{(verb)}$'

b. **ajustable** → que pot ser $\boxed{\text{ajustat}}_{(verb)}$
'adjustable → that can be $\boxed{\text{adjusted}}_{(verb)}$'

(4.5) a. que ha sofert el procés de □$_{(verb)}$(-ho/-lo/-se)
'that has undergone the process of □$_{(verb)}$(object clitics)'

b. **especialitzat** → que ha sofert el procés de $\boxed{\text{especialitzar}}_{(verb)}$(-ho/-lo/-se)
'specialised → that has undergone the process of $\boxed{\text{specialising}}_{(verb)}$(*object clitics*)'

No instructions were provided as to how to use the patterns. This decision was motivated by the time constraints set on by a web experiment, because it discourages participation to have to read too many instructions or going through too many web pages before starting. Examples were provided in the instructions, and judges did three trial adjectives (for which they were shown the expected answers) so as to clarify the task. Following standards in psycholinguistic research, no example sentences were provided for the adjectives to be examined during the experiment, so as not to bias the judges' responses.

Participants could select more than one pattern in case of polysemy. In the instructions, this concept was not mentioned, but an example was provided with some explanation. Our hypothesis was that at most two patterns would be enough to account for polysemy in our setting, because much of the polysemy occurs within two classes. As has been explained in Section 3.6.4, polysemous items (of the kind of polysemy of interest here) are usually derived adjectives with an object- or event-related meaning that take on a basic meaning, as the *econòmic* ('economical/cheap') case. However, initially, no constraint was set on the maximal number of patterns a participant could fill in, so as to test this hypothesis. The use of 3 or more patterns was strikingly unfrequent, as can be seen in Table 4.4.

#patterns	#examples	%
0	47	2%
1	1318	62%
2	460	22%
3	90	4%
4	66	3%
5	150	7%
Total	2131	100%

Table 4.4: *Number of patterns filled in when no constraints are set.*

The data in Table 4.4 correspond to responses from 85 participants. As can be seen, only in 7% of the cases were 3 or 4 patterns used. The 7% of the cases where all 5 patterns were used (as well as many of the uses of 3 or 4 patterns) correspond to participants that filled in everything, regardless of the meaning of the adjective (they even made words up). This kind of indiscriminant participant was excluded from the analysis. From the remaining cases, a manual examination revealed that they corresponded to not following the instructions (more details on participant exclusion and error review in Sections 4.2.3 and 4.2.4 below).

Therefore, we decided to change the instructions and explicitly ask participants to fill in only one or two of the patterns. This decision makes the task clearer (the ratio of indiscriminant judges decreased dramatically) and the analysis of the results easier, while not significantly decreasing descriptive accuracy. The data gathered before this decision was made were not taken into account for the analysis.

The experiment was structured as follows: [2]

- first page with introduction and classificatory questions
- second page with instructions and examples
- three training adjectives, with expected answer after the participants' response
- into the experiment: 1 page per adjective (30 adjectives)
- final "thanks" page, with a small explanation of the purposes of the experiment and the possibility for the participant to write a comment

For each participant, one of the 7 test sets of the Gold Standard was randomly chosen, and the order of the 30 adjectives to be judged was also randomised. Initially, the order of the definitional patterns was always the same (first the object pattern, then the three event patterns, then the basic pattern). We observed an overuse of the object pattern, and randomised also the order of the patterns so as to avoid ordering effects. For all analysis purposes in what follows, we only take into account responses generated with the final setup (maximum of two patterns, randomised presentation order for patterns).

4.2.3 Participants

603 participants, all self-reported as native speakers of Catalan, took part in the web experiment. Participants were recruited via e-mail to several university departments and distribution lists, and received no payment. To encourage participants to reveal their e-mail address, so that they would commit themselves with the experiment (Reips, 2002), we offered a prize of 2 vouchers of 30 euros each. The sources of participants were the following:[3]

- Friends and family
- University
  - Pompeu Fabra University: staff of the Translation and Philology Department and the Technology Department; students of Linguistics, Translation and Interpreting, Computer Science, Telecommunication Engineering, Biology, and Law.
  - University of Barcelona: professors and students of Linguistics.

[2] The experiment will be available online at `http://mutis.upf.es/˜boleda/adjectives/` for some time.

[3] Except in the case of friends and family, and in order to adhere to ethical standards, we asked for permission to advertise the experiment to the relevant authorities.

  - Professors and students of the Cognitive Science and Language Doctoral Program (from 4 Catalan universities).

- Distribution lists

  - *Carta de Lingüística* (linguistics distribution list). Scope: Catalonia.
  - *Info-Zèfir* distribution list. Audience: professionals dealing with Catalan.
  - Distribution list of the Asociación de Jóvenes Lingüistas (Young Linguists Association). Scope: Spain.

The experiment was also included in *Language Experiments*[4], a portal for psychological experiments on language, and an advert was placed in the author's homepage.

Of the 603 participants, 101 (17%) only read instructions, without classifying a single adjective. 131 (22%) filled in too few data for results to be analysed (we set the threshold at 20 adjectives). The dropout rate is quite high (39%), although we have not found reported dropout ratios for similar web experiments for comparison. Finally, 15 (2%) participants filled in 3 patterns or more for at least 20 adjectives, and were excluded for analysis purposes (these are referred to above as "indiscriminant participants"). The descriptive data in Table 4.5 correspond to the remaining 322 participants, and are all self-reported (*NR* stands for *not reported*).

Information	**Distribution**
Age	min. 14; max. 65; mean 27.5; median 23
Mother tongue	Catalan 82%; Spanish 16%; other 1%; NR 1%
Region	Catalonia 77%; Valencia 15%; Balearic Islands 4%; other 2%; NR 1%
Study level	university 89%; pre-university 8%; NR 3%
Study field	Arts 60%; Science 20%; Technical 17%; NR 4%
Knowledge in linguistics	yes 71%; no 26% NR 3%

Table 4.5: *Main characteristics of participants in web experiment.*

The prototypical participant is a university student (see median age, 23, and overwhelming university study level in Table 4.5) from Catalonia with Catalan as mother tongue. A few participants have Spanish as main mother tongue but they are also native speakers of Catalan, because of the bilingual status of Catalan society. Also, a few participants come from other regions than Catalonia, such as Valencia or the Balearic Islands[5].

Note the high reported expertise in linguistics (71% participants report themselves as having knowledge in linguistics). These data surely correspond to a wrong formulation of the question (which was "do you have knowledge in linguistics? (beyond secondary school)"), because given the age and study field of most participants, it cannot be that over 70% actually have training in linguistics. Probably, many participants answered "yes" if they know foreign languages or for other reasons. This makes it impossible to test any hypotheses about differences between participants with and without expertise in the field, which would have been a very relevant piece of data for our study.

[4] `http://www.language-experiments.org/`

[5] The question participants answered was "In which region did you grow up?".

4.2.4 Data collection and cleaning

The data were collected during March 2006. The responses were semi-automatically checked for compliance with instructions, through the procedures explained in this section.

Responses with three or more filled patterns were automatically discarded, because the instructions explicitly required judges to fill in at most two patterns. Those with more than one word were automatically identified and discarded, with the following exceptions: some clear compounds such as *ésser humà* 'human being', cases where the participant had provided more than one equivalent response, as in example (4.6a) (in these cases, only the first response was retained), or synonyms with a grading adverb, as in example (4.6b).

(4.6) a. típic → habitual, comú
typical → *habitual*, *common*

b. roent → molt calent
burning → *very hot*

Other kinds of responses with multiple words, in addition to not complying with the instructions, typically correspond to a wrong use of the pattern, as can be seen in example (4.7).

(4.7) catalanista → (que) defensa Catalunya i les seves costums i tradicions pròpies
catalanist(ic) → *(that/who) defends Catalonia and its traditions*

This participant used one of the eventive patterns to provide a gloss of the adjective *catalanista* that does not correspond to the intended use of this pattern (identifying adjectives with an event embedded in their meaning).

Responses with a part of speech that did not coincide with the one signalled in the instructions (adjective, verb or noun) were also discarded. To perform these two correction steps, a semi-automatic procedure was followed. The responses were checked against GLiCom's computational dictionary. If the right POS was not found among the readings of the word, the response was manually checked. It usually corresponded to a wrong POS, a spelling mistake, or a lexical mistake (non existing word). As for wrong POS, these responses were discarded, except for process-denoting nouns inserted in the process pattern (e.g. *americanització*, 'americanisation'). Spelling mistakes were corrected for normalisation reasons.

As for lexical mistakes, some of them correspond to interferences with Spanish (see example (4.8a)). For these cases, the equivalent in Catalan was recorded so as to normalise the data. However, most lexical mistakes were words participants invented (example (4.8b)), and these were discarded.[6] Presumably, time constraints and performance pressure where the causes of participants making words up.

(4.8) a. *mercancia* (from Spanish 'mercancía') corrected to *mercaderia* 'commodity, goods'

b. mutu → *mutuar
mutual → *? (non existing deadjectival verb*)

[6] The online version of the dictionary of the Institut d'Estudis Catalans (`http://pdl.iec.es`) was checked to ensure that the problem was not the coverage of GliCom's dictionary.

A particular kind of lexical mistake cannot be detected through this procedure: a response corresponding to another word due to a reading mistake (see example (4.9), due to a confusion with *epistològic* ‘epistological’). The cases that were detected were discarded, but as no systematic manual exploration was performed, presumably some of these mistakes remain in the data.

(4.9) epistemològic → cartes
epistemological → *letters*

Error type	**Basic**	**Event1**	**Event2**	**Event3**	**Object**	***Total***
Multiple word	6	**131**	4	3	16	*160*
Wrong POS	**92**	1	1	12	22	*128*
Non-existing word	9	6	8	11	19	*53*
Wrong reading	2	6	0	5	4	*17*
Total error	*109*	*144*	*13*	*31*	*61*	*358*
Total responses	4,388	1,304	605	504	4,341	11,142

Table 4.6: *Distribution of errors in participants’ responses.*

The distribution of the errors identified (corresponding to discarded cases) is depicted in Table 4.6. The total number of errors detected (358) corresponds to 3.2% of the data. For comparison, Corley and Scheepers (2002) excluded 3% of their experimental data in a web-based syntactic priming experiment because the prime-to-target times were too long. Our noisy data has a similar proportion. However, almost two thirds of the errors are concentrated in two cells of Table 4.6 (bold faced), which probably points to problems in the experimental design.

The first event pattern (‘that/which/who $\boxed{}_{(verb)}$’) caused 131 multiple word errors, which indicates that it was not constrained enough; the rest of the patterns were more concrete. In addition, many adjectival dictionary definitions begin with ‘que’, so that in this case the design of the experiment as dictionary definitions seems not to be optimal.

The basic pattern (‘has a meaning similar to / opposite to $\boxed{}_{(adjective)}$’) causes 92 errors where a wrong POS (mainly, a noun) was provided. There are presumably two main reasons for this high number of errors. The first reason is the large proportion of ambiguity between adjective and noun in Catalan, as discussed in Section 3.1.3 (see Table 3.2, page 22). This caused some responses corresponding to the noun homograph, not to the adjective, as in examples in 4.10[7]. The second reason is that the notion of similarity of meaning (as glossed in the definition pattern) is quite vague, so that other kinds of semantic relationships than synonymy or antonymy fit in, as can be seen in examples (4.11).

(4.10) a. obrer → patró
working-class (adjective) → *boss*

b. utilitari → cotxe
utilitary (used as noun: *utility car*) → *car*

[7] This kind of mistake corresponds both to “wrong POS” and “wrong reading”, but it was tagged as wrong POS.

(4.11) a. alegre → tristesa
joyful → sadness

b. abundant → molt
abundant → very, much, a lot

We now turn to discussing the agreement in the responses. Before analysing the results, we discuss different approaches to measuring interrater agreement.

4.3 The assessment of interrater agreement

The main factor of the data we want to analyse is the extent to which different participants agree in the classification they implicitly provide. The assessment of interrater agreement[8] (and, relatedly, reliability) is a complex area, and statisticians do not agree on a single best method or approach to address it in a variety of settings, or even within a single setting. We will restrict the discussion to the assessment of agreement with nominal categories (as opposed to ordered or continuous categories).

Many agreement indices for nominal categories have been proposed. For instance, Fleiss (1981, chapter 13) discusses 5 indices, and mentions four more. Popping (1988) (according to Lombard et al. (2002)) identified 39 different indices for nominal data. As Fleiss (1981, p. 216) put it, however, "there must be more to the measurement of interrater agreement than the arbitrary selection of an index of agreement", particularly so given that different indices provide a different impression of the reliability of the results. The following discussion is an attempt at clarifying the issues that are at stake in the assessment of agreement.

4.3.1 Overall proportion of agreement

The most straightforward measure for agreement (and the one that is most widely used) is p_o, or overall proportion of agreement (also termed *observed agreement*; Hripcsak and Heitjan (2002)). If there are two raters and two categories, their judgements can be depicted in a contingency table such as Table 4.7 (Hripcsak and Heitjan (2002, Table 1)). The categories (positive and negative) typically represent the presence or absence of a particular trait, such as a disease in medical diagnosis.

		Rater 2		
		Positive	Negative	*Total*
	Positive	a	b	a+b
Rater 1	Negative	c	d	c+d
	Total	a+c	b+d	a+b+c+d

Table 4.7: *Two-by-two contingency table.*

[8] Several terms exist for the same concept, which are somewhat field related: *intercoder agreement* is used in content analysis (Krippendorff, 1980). In NLP, the term is usually *interannotator agreement*, because most agreement measurement efforts are devoted to corpus annotation. In statistics, the preferred term is *interrater agreement*. We use the latter term because it is the most general one and because the term usually used in our field, *interannotator agreement*, is more adequate for corpus annotation than for classification of a set of lemmata.

P_o is simply the proportion of cases where judges agree in their judgement, that is, how many of the objects both judges classify as positive or negative. Accordingly, it ranges between 0 and 1. Applied to Table 4.7, the formula for p_o would be as follows Hripcsak and Heitjan (2002, p. 100):

$$p_o = \frac{a + d}{a + b + c + d} \tag{4.1}$$

If there are more than two categories, the formula can be straightforwardly extended (Fleiss (1981), Uebersax (2006)). Instead of just $a+d$, we take the diagonal of the contingency table as the cases where judges agree, as shown in Table 4.8 and Equation (4.2) (Uebersax, 2006, Table2 and Equation (4)). In this table, n_{11} represents the total number of cases where both Rater 1 and Rater 2 have assigned the object to category 1, for instance, *basic*. Cell n_{12} contains the cases where Rater 1 has assigned the object to category 1 (e.g., *basic*) and Rater 2 to category 2 (e.g., *event*). Because cases where there is agreement lie at the diagonal of the table, and cases where there is disagreement are off-diagonal, Equation (4.2) simply sums the diagonal cells (n_{ii}, or cases where indices coincide) and divides by the total number of cases (N).

		Rater 2				
		1	2	…	C	Total
	1	n_{11}	n_{12}	…	n_{1C}	$n_{1.}$
Rater 1	2	n_{21}	n_{22}	…	n_{2C}	$n_{2.}$
	…	…	…	…	…	…
	C	n_{C1}	n_{C2}	…	n_{CC}	$n_{C.}$
	Total	$n_{.1}$	$n_{.2}$	…	$n_{.C}$	N

Table 4.8: *Multi-category contingency table.*

$$p_o = \frac{1}{N} \sum_{i=1}^{C} n_{ii} \tag{4.2}$$

This formula yields an intuitive measure for interrater agreement. However, it can be artificially inflated if the categories are very unevenly distributed. Consider the case of a rare disease, for which there is an overwhelmingly large number of negative judgements, an example of which is depicted in Table 4.9 (Hripcsak and Heitjan, 2002, Table 3).

		Rater 2		
		Positive	Negative	*Total*
	Positive	4	6	10
Rater 1	Negative	8	102	110
	Total	12	108	120

Table 4.9: *Contingency table for mediocre ability to diagnose a rare disease.*

In this case,

$$p_o = \frac{4 + 102}{120} = \frac{106}{120} = 0.88$$

0.88 seems to be quite a high agreement, considering that the maximum is 1. However, it is clear from Table 4.9 that the relevant agreement is much lower, because raters disagree on 14 (out of 18) potentially positive cases. The problem is that, because most cases are negative, the poor agreement on positive cases is obscured. The raters would have high agreement in any case, just because they tag most cases as negative. Note that observed agreement would be even higher if e.g. Rater 2 were 'useless' and would always give a negative rating. In this case, p_o would be $4 + 110/120 = 0.95$.

4.3.2 Chance-corrected indices

The considerations in the previous subsection have led scholars to propose indices that correct for chance. These indices factor out the agreement that would be expected if raters would provide their judgments just randomly. The general form of the indices can be depicted as in Equation (4.3).

$$\frac{p_o - p_e}{1 - p_e} \tag{4.3}$$

In this formula, p_o is observed agreement (computed as in Equation (4.1)), and p_e agreement expected by chance. The denominator normalizes the scale so that agreement values lie at most within [-1, 1]. 1 indicates perfect agreement, 0 chance agreement (note that in this case, $p_o = p_e$), and -1 systematic disagreement (Fleiss, 1981; Carletta, 1996; Di Eugenio and Glass, 2004).

The major difference among indices is the way chance agreement is modeled, that is, what are the probabilities of each side of the dice. If the distribution of the categories is assumed to be equal (for 2 categories, $p_e = 0.5$), we end up with the S measure presented in Bennet et al. (1954, see Krippendorff (2004b)). However, this assumption is clearly wrong in many cases (as the one depicted in Table 4.9), and more refined approaches to p_e have appeared in the literature. The two most relevant approaches are Cohen's (1960) kappa (K) and Scott's (1955) pi (π). These are the two most used measures for nominal data.

The difference between Cohen's kappa and Scott's π is, as already mentioned, how they compute p_e. Cohen's kappa assumes independence of judges, and accordingly computes p_e taking into account the sum of the **product** of the marginal proportions, as Equation (4.4) shows (the equation follows the notation employed in Table 4.8).

$$p_e(K) = \frac{1}{n^2} \sum_{i=1}^{C} n_{i.}n_{.i} \tag{4.4}$$

In contrast, Scott's π assumes an equal underlying distribution of the categories across judges, that is, it assumes that the number of items in each category is the same across different judges (but not in all categories, contrary to Bennet et al.'s S). Therefore, it estimates expected agreement from the **mean** of the marginal proportions, as shown in Equation (4.3). Note that π is restricted to 2 categories. Krippendorff (1980; 2004a; 2004b) has generalised the measure to more than 2 categories in his α measure.[9]

[9] The same assumption underlies the computation of K found in Siegel and Castelan (1988). As Di Eugenio and Glass (2004, fn. 1) and Krippendorff (2004a, p. 250) note, it is an extension of π similar to α, rather than a version of kappa.

$$p_e(\pi) = \frac{1}{4n^2} \sum_{i=1}^{2} (n_{i.} + n_{.i})^2 \tag{4.5}$$

Di Eugenio and Glass (2004) argue that in dialogue modeling and other tasks related to computational linguistics, the assumption underlying π or α is more appropriate. This could also be argued for our task, because, as native speakers of Catalan, all participants should have the same (or very similar) distribution of adjectives into semantic classes. However, there is a difference between the classes themselves and the parameters given in the experiment. Indeed, different participants follow different strategies in their responses, and show different biases toward one or the other category.

Whether such a bias exists can be tested with tests of marginal homogeneity. The most standard one is the McNemar test (McNemar, 1947), which however only applies to two-way distinctions. Bishop et al. (1975) describe an alternative that allows multi-way category homogeneity comparisons. The statistic they provide, as the statistic obtained with the McNemar test, can be viewed as a chi-squared statistic with one degree of freedom, and is computed as in equation (4.6). In this equation, UD stands for upper than diagonal and LD for lower than diagonal. The statistic ignores the diagonal (elements in which raters agree) and compares the cells above the diagonal with those below the diagonal. If they are comparable, they will cancel out. If one of them is much higher than the other, this indicates a bias.

$$\chi^2 = \frac{(\sum UD - \sum LD)^2}{UD + LD} \tag{4.6}$$

In our data (the sampling methodology will be explained in Section 4.3.6), 28 out of 158 pairs of judges, that is, 18%, showed a significant bias effect ($p < 0.05$). The assumption underlying Scott's π or Krippendorff's α, therefore, does not hold in our data, so that we will use K in assessing agreement for our task. However, kappa's handling of distributional differences among raters is also not optimal, as will be discussed in the next Section. [10] For a thorough discussion of the differences between π and K (and their extensions to multi-category and multi-judge agreement computation), see Artstein and Poesio (2005b).

4.3.3 Controversy around kappa

Despite its wide use (or maybe because of its wide use), kappa is a controversial measure (see Cicchetti and Feinsten (1990); Byrt, Bishop and Carlin (1993); Uebersax (2006) and references cited therein). There are two major features of kappa that render it problematic, that have been termed *prevalence* and *bias* in the literature (Byrt, Bishop and Carlin, 1993; Hripcsak and Heitjan, 2002; Di Eugenio and Glass, 2004).

The prevalence[11] problem amounts to the fact that, for a given value of p_o, the larger the value

[10] Note, in addition, that Artstein and Poesio (2005a) formally show that increasing the number of annotators decreases the effect of bias, thus making K values more similar to α values. Artstein and Poesio (2005b, p. 21) report that in an annotation experiment with 18 subjects, they found that in a diverse range of conditions, the values of α and the extension of weighted K to multi-judge situations, β, did not differ beyond the third decimal point.

[11] This term comes from epidemiology and expresses "the ratio of the number of cases of a disease present in a statistical population at a specified time and the number of individuals in the population at that specified time" (Wikipedia; `http://en.wikipedia.org/wiki/Prevalence`).

of p_e, the lower the value of K (Di Eugenio and Glass, 2004, p. 98). This characteristic is due to the fact that if categories are very skewed, expected agreement is very high, so that no matter how large observed agreement is, it will always be severely diminished by substracting expected agreement from it.

Di Eugenio and Glass (2004) illustrate the problem with the distributions in Tables 4.10 and 4.11 (Di Eugenio and Glass, 2004, Examples 3 and 4 in Figure 3, p. 99). The categories represent “accept” or “acknowledge” codings of uses of *Okay* in English).

		Coder 2		
		Accept	Ack	*Total*
	Accept	90	5	95
Coder 1	Ack	5	0	5
	Total	95	5	100

Table 4.10: *2x2 contingency table with skewed categories.*

		Coder 2		
		Accept	Ack	*Total*
	Accept	45	5	50
Coder 1	Ack	5	45	50
	Total	50	50	100

Table 4.11: *2x2 contingency table with balanced categories.*

In both tables, p_o is 0.90. However, when data fall very predominantly in one category (in Table 4.10, $p_{Accept} = 0.95$; $p_{Ack} = .05$), K = -0.048. When the distribution of categories is balanced (in Table 4.11, $p_{Accept} = p_{Ack} = .50$), K = 0.80. This behaviour has been noted as a major flaw of K in assessing interrater agreement. However, in the case of Table 4.10, it can be argued that kappa is right in pointing out poor agreement, because there are actually no “interesting” cases for which both judges agree. They merely agree on the default case, that is, *Accept*, but they do not agree on a single *Acknowledgement* case. Note, in addition, that skewed distributions are the very problem that motivated the definition of K in the first place, as discussed in Section 4.3.2.

Some authors, like Kraemer et al. (2002, p. 2114), argue that the behaviour of kappa with unbalanced samples “merely reflects the fact that it is difficult to make clear distinctions between the [objects] in a population in which those distinctions are very rare or fine.” They claim that K = 0 “indicates either that the heterogeneity of the patients in the population is not well detected by the raters or ratings” (real disagreement), “or that the patients in the population are homogeneous” (disagreement caused by prevalence of one category). A similar point is made by Hripcsak and Heitjan (2002, p. 107): “The problem is not that kappa is too low when the sample is unbalanced. The problem is that a severely unbalanced sample does not contain sufficient information to distinguish excellent raters from mediocre ones”.

It seems, thus, that no index can account for such distributions. Byrt et al. (1993) recommend reporting a quantitative indicator of prevalence so as to be able to judge whether a low K indicates real or spurious disagreement. Cicchetti and Feinsten (1990) recommend the use of specific agreement for unbalanced samples (to be discussed in Section 4.3.4 below). As a result of the sensitivity to prevalence, chance-corrected values can only be compared across studies when the agreement data are very similar; particularly, the number of categories and their distribution. Prevalence also affects π and α.

We now turn to the bias problem. K does not assume marginal homogeneity, that is, it does not assume that the raters have a similar distribution of objects into categories. However, having different distributions of categories implies having fewer chance agreements, and having fewer chance agreements (lower p_e) has the effect of increasing K. This leads to the paradox that, for the same observed agreement, K is higher for raters with a dissimilar distribution of objects into categories than for raters with a similar distribution. Krippendorff (2004a, p. 246) says that p_e as defined in K "is the agreement that can be expected when the two observers' proclivity to use their categories differently is assumed and taken for granted".

We reproduce Di Eugenio and Glass (2004)'s illustration of the bias problem with the distributions in Tables 4.12 and 4.13 (Examples 5 and 6 in Di Eugenio and Glass (2004, Figure 4, p. 99)).

		Coder 2		
		Accept	Ack	*Total*
	Accept	40	15	55
Coder 1	Ack	20	25	45
	Total	60	40	100

Table 4.12: *2x2 contingency table with similar marginal proportions.*

		Coder 2		
		Accept	Ack	*Total*
	Accept	40	35	75
Coder 1	Ack	0	25	25
	Total	40	60	100

Table 4.13: *2x2 contingency table with very different marginal proportions.*

In both tables, p_o is 0.65. However, when the marginal proportions are similar (Table 4.12), K = 0.27. When they are dissimilar (Table 4.13), counter intuitively, K is higher (0.42), so that it indicates that the agreement is higher. Krippendorff (2004a, p. 247) notes, for a similar example, that the "mismatches, initially populating both off-diagonal triangles, have now become unevenly distributed, occupying only one. What has increased thereby is not agreement but the predictability of the categories used by one coder from the categories used by the other. ... predictability has nothing to do with reliability."

However, Artstein and Poesio (2005b, p. 19) argue that "the bias problem is less paradoxical than it sounds". They note that observed agreement and expected agreement are not independent (they are computed from the same data), so that "if it so happens that two data sets have similar observed agreement and different biases ... then the data set with the higher bias is indeed more reliable. A high bias may indicate that the coding process is defective, but it also indicates that whatever data are agreed upon are less likely to be the result of chance errors."

There is a further substantial source of controversy around K, which however arises for any numerical index of agreement: how to interpret its value. A value of zero indicates mere chance agreement, and a value of one, perfect agreement; but what about 0.8? And 0.6? Where can we draw the line between an acceptable value for a Gold Standard? Krippendorff (1980) notes that the acceptable value of an index of agreement depends on the kind of decision that has to be made with the agreement data. If human lives depend on the result (as for instance with medical diagnosis), a higher score will be demanded. In the academic context, several scales have been proposed.

Landis and Koch (1977) propose a 6-way division for the strength of evidence for agreement: <0 (poor), 0-0.20 (slight), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), and 0.81-1.00 (perfect). Fleiss (1981, p. 218) establishes a coarser division, distinguishing between poor ($<$0.40), fair to good (0.40-0.75), and excellent ($>$ 0.75) agreement beyond chance.

Carletta (1996) suggests adapting Krippendorff's (1980) scale for Content Analysis to Computational Linguistics. Krippendorff requires values over 0.8 for data to be deemed reliable, and poses that values over 0.67 allow "tentative conclusions to be drawn". It has to be noted that Krippendorff proposes this scale for α, not for K. However, even if α consistently yields lower values than K (Artstein and Poesio, 2005b, among others), Krippendorff's scale is stricter than the other two.

Artstein and Poesio (2005b, p. 28) state that in their research in assessing semantic judgments they have found "that substantial, but by no means perfect, agreement among coders resulted in values of κ or α around the .7 level. But we also found that, in general, only values above .8 ensured a reasonable quality annotation". They also note that in many tasks (for instance discourse annotation), even a lower threshold is difficult or impossible to achieve.

To sum up, there is little doubt that an agreement value exceeding 0.8 can be considered to be valid for academic purposes, but there is wide disagreement as to the meaning of values below that. [12]

4.3.4 Per category agreement

All measures discussed so far are overall measures, that is, they do not provide information on agreement for a single category. A useful descriptive measure to address category agreement is proportion of specific agreement, or p_s (Fleiss, 1981; Hripcsak and Heitjan, 2002; Uebersax, 2006). For dichotomous data, this measure distinguishes between agreement for positive cases (p_{sPos}) and agreement for negative cases (p_{sNeg}). These measures are computed as in Equations (4.7) and (4.8), following the nomenclature in Table 4.7.

$$p_{sPos} = \frac{2a}{2a+b+c} \tag{4.7}$$

$$p_{sNeg} = \frac{2d}{2d+b+c} \tag{4.8}$$

These proportions correspond to Dice coefficients for positive and negative ratings respectively (Fleiss, 1981). According to Uebersax (2006), specific agreement addresses the objection raised against observed agreement, namely that high values can be obtained by chance alone. If both p_{sPos} and p_{sNeg} are high, it can be concluded that the observed level of agreement is higher than the level that would be obtained by chance alone. In the example of Table 4.9, the value for p_{sNeg} is 0.94, suggesting high agreement. However, the value for p_{sPos} is 0.36, showing that there is a much lower agreement for the positive cases, presumably the most important ones in the decisions to be made based on the human judgments.

[12] The problems discussed in this Section and other considerations lead Uebersax (2006) to a radical conclusion, namely, that κ can only be used to determine whether the agreement among two given judges exceeds chance agreement. This is clearly not useful for the present purposes, for it is the *degree* of agreement that is at stake here.

An alternative is to use kappa on a per-category basis, which Fleiss (1981) argues for. He shows that the per-category values of kappa remain stable with 5 different indices of agreement (including p_o and p_s). To compute kappa on a per-category basis with multiple categories (as is our case), each of the distinctions is lumped into a Yes/No, 2-category distinction, and kappa is computed as if it were a dichotomous variable.[13] In our case, the distinctions would be Basic/Not basic, Event/Not event, etc.

In this case, the distributions will surely be skewed (positive cases being fewer than negative cases for most categories), so that values of per-category kappa will be lower than the overall kappa in most cases. The point of a per-category agreement value is to compare across categories, so this is not a problem in using it. The problem is rather that if one category is much smaller than another one, the actual agreement could be obscured due to the prevalence problem.

The same methodology can be applied to measure p_s for more than 2 categories. In this case, each category's p_s is the p_{sPos} it would have if the distinction were dichotomous Uebersax (2006). The computation of p_s for each category with multiple categories is shown in Equation (4.9) (following the nomenclature in Table 4.8).

$$p_s = \frac{2n_{ii}}{n_{i.} + n_{.i}} \tag{4.9}$$

4.3.5 Estimation of standard error

No matter which agreement measures are used, and at what level (overall or category-specific) they always correspond to the estimation of the population parameters, the “real” agreement values in the population. They are estimated from a sample (the set of coders and the set of objects coded), and thus are subject to sampling error. Therefore, not only the agreement values, but the standard error or confidence interval should be reported. This issue is generally not tackled in the Computational Linguistics literature. As an example of its relevance, consider a fictitious example, in which a K value of 0.6 is obtained. The reliability of this estimate (how well it approximates the agreement value for the task) is very different if the confidence interval for these data ranges from 0.55 to 0.65 than if it ranges from 0.30 to 0.90. In the latter case, the actual value could be indicative of of poor, fair to good, or excellent agreement, using the scale proposed by Fleiss (1981, p. 218; see Section 4.3.3), so that its results really would be meaningless to assess the reliability of the classification process.

A practical problem is the controversy and complexity of the estimation of standard error for K, particularly when there are more than 2 categories, as is our case. For the 2-judge, 2-category case, Fleiss (1981, p. 221) provides a formula for standard error and interval confidence computation, according to the underlying hypothesis that K corresponds to a value other than 0 (formulas for the $K = 0$ hypothesis are provided earlier in the chapter). The formula for the multi-judge, 2-category case is provided in Fleiss (1981, p. 228). However, the underlying assumption is that judgments are independent and therefore $K = 0$, an unappropriate assumption for our field, and probably for any agreement measurement purpose, because we do not want to

[13] Note that in this setting it would also be possible to compute p_o for each category. However, this approach is subject to the same criticism p_o is generally subject to: if one of the categories is substantially larger than the other one, the actual agreement is obscured. Indeed, when lumping distinctions together, it is more probable that the negative category be much larger, so that the values obtained are of no analytical use. The use of kappa makes it possible to overcome this limitation, because of its factoring out chance agreement.

analyse whether agreement is better than chance (it will presumably always be), but to assess the degree of agreement.

Finally, for the multi-judge, multi-category case, Fleiss does not provide the standard error except for the case when there is no missing data (and even for this case, the underlying hypothesis is still $K = 0$). Recently, e.g. Lui et al. (1999) and Altaye et al. (2001) have proposed two different approaches to the computation of the confidence interval for the multi-class problem, each relying on different assumptions, and both mathematically very complex.

Bootstrap resampling, a statistical estimation technique, has been proposed for the estimation of confidence intervals for agreement measures (Krippendorff (2004a, pp. 237-238) applies it to α and Lee and Fung (1993) to K). With this technique, a large number of subsamples from the data are drawn and the statistic of interest is computed for each subsample. The statistics thus collected form a probability distribution, from which the confidence interval can be computed. However, Krippendorff does not describe the implementation of the bootstrap resampling procedure in detail (does he resample only over judges or also over objects?), nor the model used to estimate the confidence interval.

Typically, agreement scores are used with a relatively large number of objects to be classified and a small number of judges to classify them, and the proposals in the literature for standard error computation are adapted to this setting. Our situation is the reverse: we have a large number of judges for each object (32 to 59 depending on the adjective) and a small number of objects per judge (about 30). All in all, we have data for over 7,000 pairs of judges, each yielding an agreement score. Given that we have so many estimates of the actual agreement for our task, we could compute a confidence interval in the usual way, using the t-distribution. However, this would not be adequate, because each judge participates in n pairs of judgments, n being the number of judges for a particular test set. Thus, if we had 100 judges, there would be only 100 independent events, but 4,950 agreement scores. In general, there are $n(n-1)/2$ pairs for a given number of judges n.

An alternative approach to statistically robust confidence interval estimation is to estimate pair-wise agreement instead of multi-judge agreement. In the approach we propose, judges are randomly split into pairs, so that the unit of analysis is the judge *pair*. The agreement values for pairs of judges form a distribution with independent values, so that the confidence interval can be estimated using the t-distribution. The solution is not optimal in that it does not integrate different behaviours of judges, only compares each judge with another one, randomly chosen. Also note that this kind of solution needs quite a lot of judges per object, so it is not applicable in many research situations. However, it responds to the usual practice (in medicine and other fields) of reporting mean kappa values when multiple judges are involved, as mean pair-wise values are an approximation of multi-judge behaviour.

4.3.6 Agreement measurement in Computational Linguistics

Interrater agreement is not much discussed in Computational Linguistics or NLP. Many resources used in machine learning experiments for POS-tagging, parsing, or lexical semantics, have been annotated by only one annotator, or more accurately, by several annotators working on different portions of the resource. The discussion of interrater agreement is only tackled (if ever) when describing the resources: once a resource is created, it is accepted as a Gold Standard and its reliability is usually not discussed in research using it for machine learning or information extraction purposes.

For “classical” resources a decade ago, agreement was usually only measured with p_o. For instance, within WordNet an experiment was performed in which naive and lexicographer judges tagged a text with WordNet senses (Fellbaum et al., 1998). Agreement is discussed in terms of percentage of observed agreement. As for the Penn TreeBank, interrater agreement is discussed mainly in the context of tagging from scratch vs. correction of proposals made by an automatic tagger (Marcus et al., 1993). The authors report mean interrater disagreement, which is equivalent to mean observed agreement: $p_o = 1 - d_o$ (d_o indicating observed disagreement).

Brants (2000) introduces standard metrics in machine learning, namely accuracy, precision, and recall, for the evaluation of interrater agreement in the annotation of a German corpus (part of speech and syntactic annotation). These metrics were justified by the fact that he compared two independent annotations with the final, consensuated annotation, acting as a Gold Standard. The initial annotations were not consistently produced by the same pair of judges, but by 6 different annotators, so that Brants (2000) claims to report on “the overall agreement of annotations, averaging over different “styles” of the annotators, and averaging over annotators that match very well or very poorly”. This is against standard practice, in which annotations provided by the same pair of judges are compared. Vilain et al. (1995) had also used recall and precision so as to account for partial matches in coreference annotation.

The use of evaluation metrics is problematic in that they implicitly assume a valid Gold Standard. If the annotation task is precisely directed at creating a Gold Standard, it is somewhat circular to assess the degree of agreement in comparison with an already given Gold Standard. The criticism against descriptive measures such as p_o is also valid for evaluation metrics: they do not take into account chance agreement.

Chance-corrected measures have been discussed mostly in the setting of more controversial tasks, namely, dialog act tagging (Carletta, 1996; Di Eugenio and Glass, 2004), anaphoric relations (Passonneau, 2004; Artstein and Poesio, 2005b), or word sense disambiguation (Véronis (1998)). The kappa measure was introduced in these areas, and it has become a standard for agreement measurement in NLP, most notably due to the influential squib by Carletta (1996). Di Eugenio and Glass (2004) discuss some of the problems concerning kappa introduced in Section 4.3.3, namely prevalence and bias, and the alternative measurement of p_e using the product (K) or the mean (π, α). The use of α has also been recently explored in Passonneau (2004), Artstein and Poesio (2005b), and Poesio and Artstein (2005).

One of the aspects that is more challenging in NLP tasks, particularly in lexical semantics, is the assessment of agreement when multiple categories are allowed, as is the case with polysemy judgments. Recall that we allowed judges to select more than one definitional pattern, that is, to assign lemmata to more than one class. Véronis (1998) deals with this situation, and he proposes separately computing full agreement and overlapping agreement.

In computing full agreement, two judgments are considered to agree if and only if all classes assigned coincide. For instance, an assignment to basic and another to basic and event would count as a disagreement. This is unsatisfactory, because it can be presumed that in case of polysemy, in many cases one judge will only record one sense because the other one is not salient enough at the time of the experiment (depending, among other factors, on the judge’s general use of Catalan and the linguistic interactions previous to the experiment). Another reason is somewhat the reverse: a particular, monosemous adjective, can be classified into two classes (two patterns) because of the judge paraphrasing the same sense with two patterns. In these cases, considering there to be a disagreement is simply wrong.

For that reason, Véronis (1998) proposed taking *overlapping agreement* (which he called *Max*) into account. Under this definition of agreement, an assignment to basic and another to basic and event count as an agreement, that is, two judgments agree if at least one of the assignments coincides. Véronis (1998) only estimates overlapping agreement in terms of proportion of observed agreement, not in terms of kappa. We will introduce a natural way to estimate kappa values for overlapping agreement shortly below.

Overlapping agreement is over-indulging to agreement mismatches, because it assigns equal weight to full and partial evidence for agreement. What we need is measure that assigns different scores (weights) to different types of agreement. Passonneau (2004) and Artstein and Poesio (2005b) also argue for the neeed of taking weighted measures into account for many Computational Linguistics tasks. They focus on a particular task, namely, anaphoric relation annotation, although Artstein and Poesio (2005b) claim that similar considerations can be made for many other tasks in Computational Linguistics, such as discourse deixis annotation, summarization, segmentation, or word sense tagging.

One such measure is weighted kappa (Cohen, 1968), the most popular extension proposed for kappa. It accounts for cases where “the relative seriousness of each possible disagreement [can] be quantified” (Fleiss, 1981, p. 223). In this computation, agreement weights can be assigned to each cell of the two-by-two contingency table. Consider weights w_{ij}, where $i = 1, \ldots, C$ and $j = 1, \ldots, C$ when computing C-category agreement. The weights are subject to the following constraints (Fleiss, 1981, p. 223; equations 13.24 to 13.26):

1. $0 \leq w_{ij} \leq 1$ (weights range between 0 and 1)
2. $w_{ii} = 1$ (exact agreement is given maximal weight)
3. $0 \leq w_{ij} < 1$ for $i \neq j$ (disagreements are given less than maximal weight)
4. $w_{ij} = w_{ji}$ (the two raters are considered symmetrical; a particular kind of agreement is given the same weight independently of its position in the table)

We can then define wp_o, wp_e and wK (weighted proportion of agreement, weighted expected agreement, and weighted kappa) as follows (Fleiss, 1981, p. 223; equations 13:27 to 13.29):

$$wp_o = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{C} \; w_{ij} \, n_{ij} \tag{4.10}$$

$$wp_e = \frac{1}{N^2} \sum_{i=1}^{C} \sum_{j=1}^{C} \; w_{ij} n_{i.} \, n_{.j} \tag{4.11}$$

$$wK = \frac{wp_o - wp_e}{1 - wp_e} \tag{4.12}$$

These formulas are equivalent to their unweighted versions (Equations ((4.2)-(4.4))), except for the fact that all cells are considered (instead of only the diagonal) and can potentially add some value to the final score. Fleiss and other researchers note that the standard kappa is a particular case of the weighted kappa, the case where $w_{ij} = 0$ for all $i \neq j$.

Weighted kappa discriminates between different kinds of agreement: disagreements, partial agreements, and agreements. In this way, we can account for partial agreement in a principled manner, without artificially inflating agreement scores as with overlapping agreement.

Weighted kappa was primarily designed for ordered scales where no numeric interpretation is possible (e.g. to judge a policy as very bad, bad, good, or very good), so as to give more weight to disagreements such as “bad” vs. “good” than to “good” vs. “very good”. It is not clear how to objectively establish a weight that accounts for these situations. In general, the weighting scheme is difficult to justify on independent grounds, which adds to the difficulty in interpreting the values of kappa, because values vary a lot depending on the weighting schema used (Artstein and Poesio, 2005b).

In our setting, the definition of partial agreement is clear: there is partial agreement when there is some overlapping (but not coincidence) between the classes assigned to a lemma by two judges. Which is the appropriate weight for these cases? It is possible to give a quite principled answer if we model the way the decisions are made. It can be argued that judges make three independent decisions: whether an adjective is basic or not, whether it is event or not, and whether it is object or not.[14] Agreement on each of the decisions can be assigned equal weight, 1/3, and thus we can model partial agreement. If all decisions are made in the same direction (e.g., basic/basic, or basic-object/basic-object), we assign full weight, that is, one. If a judge assigns only one class and another two classes with some overlapping (e.g., basic/basic-object), they have made the same decisions for two out of the three questions, so we assign a weight of 2/3. Finally, if both judges assign two classes but there is only one overlapping (e.g., basic-object/event-object), they have only made the same decision for one of the questions, so we assign a weight of 1/3.

A weakness of this approach is that, if strictly applied, it would imply assigning a weight of 1/3 to monosemous non-agreement (e.g., a judge assigns basic and another one event), because one of the decisions (not object) has been made in the same direction. We will assign 0 to this situation by placing a further restriction on the weighting scheme, namely, that for weight w_{ij} to be > 0, there has to be at least one positive agreement.

Table 4.14 depicts the proposed weights for our classification task. The weighting schema and the reasoning behind it can be generalised to other tasks involving polysemy.

		Rater 2					
		B	BE	BO	E	EO	O
Rater 1	B	1	2/3	2/3	0	0	0
	BE	2/3	1	1/3	2/3	1/3	0
	BO	2/3	1/3	1	0	1/3	2/3
	E	0	2/3	0	1	2/3	0
	EO	0	1/3	1/3	2/3	1	2/3
	O	0	0	2/3	0	2/3	1

Table 4.14: *Weights to account for agreement with polysemous assignments.*

Weighted kappa also offers a natural way to accomodate the notion of overlapping agreement (Véronis’s *Max*), namely, to assign a weight of 1 to all cells where there is some overlap between the categories involved. To compute overlapping agreement, all non-zero cells in Table 4.14 would contain a 1. This serves to estimate an upper bound for agreement, being maximally

[14] In fact, the decisions are not completely independent, because of the constraint we have set to a maximum of 2 responses.

indulgent as to agreement mismatches. Note that this representation violates Fleiss's constraint number 3 stated in page 84 above. However, because we use it to establish an upper bound, not as an actual agreement measure, this violation is warranted.

4.3.7 Summary

To sum up, in general we have found that descriptive measures such as the proportion of overall agreement, p_o, and proportion of specific agreement, p_s, are useful in agreement studies because they provide a common sense value for agreement measurement. In addition, p_s can point to difficulties with a particular category, or with positive or negative judgments for a dichotomous decision.

Chance-corrected measures such as the kappa coefficients are also useful to factor out the agreement that would be expected by chance alone. However, these coefficients have well-known properties (sensitivity to prevalence and bias) that make it advisable to use them with care. Many authors also raise concerns about the interpretation of agreement values, although there seems to be some consensus on the fact that values above 0.75 indicate good agreement.

Reporting category specific agreement values in addition to overall agreement values facilitates the identification of difficult distinctions and is particularly useful for studies with a small number of categories, such as our own experiment.

For many tasks in Computational Linguistics, the notion of degree of agreement (or, equivalently, disagreement) is needed: identification of members of an anaphoric chain, polysemy judgments, segmentation, etc. Weighted measures of agreement, such as wK and α, are natural choices to represent differing degrees of agreement. However, the use of a weighting scheme adds a further parameter that makes the interpretation of the values even more difficult, because values vary a lot depending on the scheme chosen. We propose a weighting scheme that responds to an explicit model of the task that judges face. However, other schemes could be proposed. For that reason, the inclusion of full agreement (only an exact assignment counts as an agreement) and overlapping agreement (any overlap in the classification counts as an agreement) values provides useful information, namely, the upper and lower bounds for the agreement values.

An issue that is not often tackled in the literature concerning agreement in different fields, and particularly in Computational Linguistics, is the computation of confidence intervals rather than a single agreement value. This makes it impossible to distinguish cases where agreement varies greatly from one to another pair of annotators, or from one to another subset of the data, from cases where values are relatively stable. In the latter case, the agreement estimate is more reliable than in the first case. However, the computation of the standard error for chance-corrected indices is a difficult and as yet unresolved issue.

We have proposed a simple, robust approach that involves randomly pairing judges and computing the confidence interval using the standard t-statistic. This simple solution comes at the price of not performing multi-judge comparisons, and also is only applicable when a large number of judges is involved in the annotation process.

From the discussion, it follows that reporting a single (or even multiple) index of agreement is not enough to achieve an understanding of the agreement patterns in the data and the sources of disagreement. In what follows, we will explore some pieces of data that provide further insight.

Finally, it has to be noted that alternatives to chance-corrected indices for the assessment of interrater agreement have been proposed. Some of them are briefly presented in Artstein and Poesio (2005a) –see also Uebersax (2006).

We now turn to discussing the agreement results for our web experiment.

4.4 Results

For all analysis purposes, we consider the three verbal definitional patterns presented in Section 4.2.2 as indicative of a unique class, namely, the event class. Together with the constraint to a maximum of two responses, this modelling gives 6 possible responses or classes for a given lemma, divided into the following two subkinds of classes:

1. monosemous classes: basic (B), event-related (E), object-related (O).
2. polysemous classes: basic-event (BE), basic-object (BO), event-object (EO).

Recall from the Section 4 that agreement scores were obtained in randomly pairing the judges for each test set. The available number of judge pairs per test set ranges between 19 and 29, as shown in Table 4.15. [15]

Set 1	**Set 2**	**Set 3**	**Set 4**	**Set 5**	**Set 6**	**Set 7**	***Total***
21	21	21	19	29	23	24	158

Table 4.15: *Judge pairs per test set.*

Following the discussion in Section 4.3, we compute agreement scores using three different agreement definitions: full agreement (all class assignments coincide, including polysemous judgments), weighted agreement (different kinds of agreements add different numerical values to the score), and overlapping agreement (if at least one class assignment coincides, judges are considered to agree). For each of the definitions, observed agreement (p_o, wp_o for weighted p_o, and op_o for overlapping p_o) and kappa (K, wK, and oK) values are reported.

The estimates for the agreement values were obtained as follows. For each test set, the mean and standard deviation of the agreement scores were computed. The standard error for the mean of each test set was obtained using the t distribution. The full results are included in Appendix A (Table A.1). A summary over all test sets is reported in Table 4.16.

In Table 4.16, mean and standard deviation (SD) of the 7 mean agreement scores are shown in the first column. The second column contains the mean and standard deviation of the standard error values (SE). The mean and standard error values in Table 4.16 are averages over all test sets. The confidence interval depicted in the last column of Table 4.16 specifies the (average) expected range of values for the population agreement. It is obtained by summing and substracting the standard error from the mean.

Table 4.16 shows that p_o values for our task are between 0.37 and 0.51, and K values are between 0.20 and 0.34. These values represent a very low level of interrater agreement. Recall

[15] The total is 158 pairs, corresponding to 316 judges. For 6 out of the 7 test sets, an odd number of judges was obtained; for these cases, one of the judges was randomly discarded.

Agr. def.	**Est.**	**Mean±SD**	**SE±SD**	**Conf. int.**
full	p_o	0.44±0.03	0.07±0.01	0.37-0.51
	K	0.27±0.04	0.07±0.01	0.20-0.34
partial	wp_o	0.66±0.02	0.04±0.01	0.62-0.70
	wK	0.38±0.04	0.07±0.02	0.31-0.45
overlapping	op_o	0.78±0.02	0.05±0.01	0.73-0.83
	oK	0.51±0.05	0.09±0.02	0.42-0.60

Table 4.16: *Overall agreement values.*

that p_o and K assess agreement in the most strict definition (*full*). At the other end, op_o and oK are overly lax: they count all overlapping assignments as agreement. According to Table 4.16, op_o ranges between 0.73 and 0.83, and oK between 0.42 and 0.60.

We have argued that weighted measures of agreement allow us to be neither too strict nor too lax, by assigning different scores to full agreement, different types of partial agreement, and disagreement. In Section 4.3.6, we have motivated our weighting scheme on independent grounds; however, the establishment of a weighting scheme is always subjective. This is why it is useful to provide the other two measures, as lower and upper bounds for the agreement values. Weighted observed agreement (wp_o) ranges from 0.62 to 0.70, and weighted kappa (wK) from 0.31 to 0.45. These values are between the p_o/K and op_o/oK values, as expected. The weighting scheme in Table 4.14 accounts for partial agreement in a sensible manner.

From the discussion it follows that the kappa value for our task is higher than 0.20 (lower extreme of the confidence interval for K) and lower than 0.60 (upper extreme of the confidence interval for oK). We consider the best estimate to correspond to wK, so that the kappa of the web experiment ranges from 0.31 to 0.45.

This range is very low, too low for the data to be considered reliable for academic purposes. Recall that Krippendorff (1980) demands as a very minimum a 0.67 value for his α measure (which yields slightly lower values than K). In the interpretation in Fleiss (1981), these values represent poor to fair agreement. Landis and Koch (1977) would rather consider them to be fair to moderate.

In studies having to do with the semantics of natural languages, high agreement values are very difficult to obtain. In Poesio and Artstein (2005), an experiment in which 18 linguistic students tag anaphoric relations is analysed. The authors report K values around 0.63-0.66. They also note that if a trivial category is dropped ("place", defined as an explicitly named set of five British railway stations), K drops to 0.45-0.50.

In Merlo and Stevenson (2001), the automatic classification of verbs into unergative, unaccusative and object-drop is discussed. Three judges with a high level of expertise tagged 59 different verbs. Despite the expertise of the judges, their kappa scores range between 0.53 and 0.66 (p_o 0.70 to 0.77).

Véronis (1998) reports on experiments on sense tagging French words with the goal of building a Gold Standard for the SensEval competition. Six students of linguistics with no training in lexicography tagged 60 highly polysemous words (20 adjectives, 20 nouns and 20 verbs) with the set of senses listed in the *Petit Larousse* dictionary. The resulting pairwise agreement was around 69% and weighted kappa around 0.43 (0.41 for adjectives and verbs, 0.46 for nouns).

B	BE	BO	E	EO	O
0.30	0.06	0.13	0.14	0.06	0.30

Table 4.17: *Class frequency in participant's data.*

[16] On a previous task involving the decision of whether a given word is polysemous or not in a set of contexts, "Full agreement on polysemy was achieved on only 4.5% of the words."

All these values are well below the ideal 0.8 threshold for kappa, which can be taken to indicate that the field of computational semantics is not mature enough to generally yield reliable classifications. However, most of the values reported are higher than our 0.31-0.45 values. While the figures are not entirely comparable (parameters such as the number and distribution of the classes and the evaluation procedures differ from the studies cited to the one presented here), they indicate that the agreement we obtain is lower than that for other semantic tasks.

We provide three main explanations for the low level of agreement. A technical explanation is that the distribution of classes is quite skewed, which, due to the prevalence problem, makes kappa values rapidly diminish when deviating from perfect agreement. Table 4.17 shows that on average participants assign much fewer items to polysemous classes (BE, BO, EO) than to monosemous classes. The basic and object classes are most frequent.

A second explanation is the fact that naive subjects were used. For the second Gold Standard explained in Section 4.1, in which 3 expert judges tagged 80 lemmata using the same classes as in the web experiment, interrater K values ranged between 0.68 and 0.80. These values are only comparable to our range for overlapping kappa, because judgments about polysemy were ignored for this Gold Standard. They are substantially higher than the oK estimate for the web experiment (0.42-0.60). This suggests that a high level of expertise is required for our task, which however would prevent large scale experiments as the one presented in this chapter.

Finally, an alternative (or complementary) explanation is that the design of the task could be unclear, something compatible with the high dropout rate. Even after the series of tests and subsequent refining of the experimental design explained in Section 4.2, some participants expressed doubts about the task. The analysis of the results provided in the rest of the chapter makes it clear that in many cases the judges' responses did not correspond to the intended uses of the definitional patterns. The sources of confusion will be explored in Section 4.5.

So as to check whether some classes contribute more to disagreement than others (which could indicate difficult or unclear distinctions), we report per-class agreement values in Table 4.18. For comparison, the second column shows the relative frequency of each class (data in Table 4.17). The detailed results per test set are included in Appendix A (Table A.2).

As in Table 4.16, values are obtained by averaging the mean and standard error values for each test set. The relevant measures for class-specific agreement are proportion of specific agreement (p_s) and K. Recall from Section 4.3.4 that p_s serves for dichotomous decisions (class X versus not class X) and ignores agreement on negative cases. Because it does not make sense to compute weighted per-class agreement, weighted measures such as wK and oK have not been computed.

[16] The computation of weighted K used in Véronis (1998) is not clear to us. He defines a weighted percentage agreement measure using the Dice coefficient, and then states that "In order to account for partial agreement, k was computed on the weighted pairwise measure using the extension proposed in Cohen (1968).".

Cl.	**Freq.**	**Est.**	**Mean±SD**	**SE±SD**	**Conf. int.**
B	0.30	p_s	0.48±0.06	0.09±0.02	0.39-0.57
		K	0.29±0.04	0.1±0.02	0.28-0.29
BE	0.06	p_s	0.13±0.03	0.11±0.02	0.02-0.24
		K	0.09±0.03	0.1±0.02	-0.01-0.19
BO	0.13	p_s	0.15±0.04	0.1±0.02	0.05-0.25
		K	0.08±0.05	0.09±0.02	-0.01-0.17
E	0.14	p_s	0.29±0.08	0.12±0.02	0.17-0.41
		K	0.22±0.07	0.12±0.02	0.10-0.34
EO	0.06	p_s	0.13±0.06	0.12±0.04	0.01-0.25
		K	0.09±0.06	0.11±0.04	-0.02-0.20
O	0.30	p_s	0.52±0.18	0.08±0.02	0.44-0.60
		K	0.36±0.16	0.09±0.02	0.27-0.45

Table 4.18: *Class-specific agreement values.*

	ρ	t	**df**	**p**
p_s-*freq.*	0.97	8.7	4	0.0009
K-*freq.*	0.93	4.9	4	0.008

Table 4.19: *Correlation between agreement measures and class frequency.*

In Table 4.18, the lowest agreement values are obtained for polysemous categories. In fact, because the confidence interval includes zero for classes BE, BO, and EO, K values can not be considered to exceed chance for polysemy judgments. As for monosemous classes, K values are highest for the object class (0.27 to 0.45) and the basic class (0.28-0.29), and lowest for the event class (0.10-0.34).

Note, however, that the prevalence problem discussed in Section 4.3.3 seems to affect both p_s and K values: classes assigned to few items tend to have low p_s and K values.

In fact, the correlation between the proportion of cases assigned to a class (column *freq.* in Table 4.18) and both agreement values is very high and, despite the few degrees of freedom, statistically significant. The result of a two-sided, paired-sample correlation test is shown in Table 4.19. The high correlation values (0.97 and 0.93) could indicate that the p_s and K values are strongly biased by class frequency. The question is whether there is really lower disagreement for the classes with lowest frequency, that is, whether p_s and K indicate real or spurious disagreement introduced by frequency bias.

We rather think that the results in Table 4.18 correspond to real agreement patterns. It makes sense that polysemy assignments be least consistent. Also, in the next section we will see that there are good reasons to think that distinctions involving the event class are confusing, supporting the lower agreement values for the event class, as opposed to the basic and object class. In our case, thus, less frequent classes seem to give rise to more disagreements than more frequent ones.

4.5 Exploring the sources of disagreement

It is clear from Section 4.4 that agreement is too low for the results to be valid for Gold Standard building purposes. One possibility would be to discard objects for which agreement is particularly low, as is done in McCarthy (2001) for verb judgments regarding alternations or Mihalcea et al. (2004) in building a Gold Standard for word sense disambiguation in Romanian. Krippendorff (1980) strongly argues against this type of solution, because it biases the experimental material. He argues that the Gold Standard building procedure should be reworked upon until reliable results are obtained.

We have pursued a further alternative, which is to let experts collectively classify the Gold Standard. This Section explains the methodology used to build the final Gold Standard, the differences between a classification based on participants' data and one based on expert data, and further analysis possibilities opened by the use of multiple judges to annotate the data.

4.5.1 Expert Gold Standard and participant's classification

An expert Gold Standard was built for two purposes: the first one, to compare with the data from the participants, so as to be able to detect systematic biases or problems in the design of the web experiment. The second one, to have a basis for comparison and analysis when performing the machine learning experiments explained in Chapter 6.

Three experts in lexical semantics (the author of the thesis, one of the supervisors, and a researcher pursuing a PhD on Catalan adjectives) gathered 3 times, each time for a 2-hour session. They reviewed each of the 210 adjectives in the Gold Standard, and assigned them to one or two classes on the basis of several pieces of information: their own intuitions, a Catalan dictionary (Institut d'Estudis Catalans, 1997), corpus examples, and the data from the participants. Decisions were reached by consensus, so as to avoid individual biases as far as possible.

Note, however, that the expert Gold Standard thus built is not more reliable than the data from the web experiment. For reliability, reproducibility is a necessary condition, and the methodology used for the expert Gold Standard does not allow assessment of reproducibility. However, it does provide a good indication of the kind of classification that experts in the field (as opposed to naive native speakers) would build for the given set of items.

In the course of building the Gold Standard, some systematic problems with the classification arose. An example are ideology-related adjectives, such as *comunista*, *anarquista*, *feminista*. As discussed in Section 3.6, these adjectives share the difficulties posed by nationality-related adjectives, namely, they seem to be underspecified between a basic and an object reading. These adjectives were systematically tagged as polysemous between basic and object.

Another example are adjectives that do not fit into the classification because they seem to participate in more than one class, without being polysemous. For instance, *fangós* ('muddy') is defined in a Catalan dictionary (Institut d'Estudis Catalans, 1997) as "full of mud". It bears a clear relationship to the object *mud*, but it is not the "related to" relationship typical of object adjectives. The semantics of the adjective is influented by the semantics of the noun, which is a mass noun, so that the resulting adjective has some properties of basic adjectives. It was coded as object (monosemous classification) by the experts due to the fact that the underlying object is quite transparent.

This example contrasts with adjective *nocturn* ('nocturnal'), which also bears a clear relationship to the object *night*. In some cases, the relationship can be viewed as the empty "related to" relationship, as in *animal / lluna / hores / tren nocturn(a/es)* ('nocturnal/night animal / moon / hour / train'). However, in most cases, the adjective has a more specific meaning, which is better paraphrased as "that happens at night", as in *vida / espectacle / passeig / cant nocturn* ('night life / show / stroll / song'). All these nouns are event-denoting nouns. Because these two senses are quite differentiated, the adjective was classified as polysemous between a basic and an object reading.

Finally, note that our classification and the explanation of polysemy provided in Section 3.6.4 assumed mainly polysemy between basic and object-related readings, on the one hand, and basic and event-related readings, on the other hand. However, the inspection of the Gold Standard made it clear that in some cases, event and object polysemy was present in the data. For instance, adjective *docent* has an event reading, as in examples in (4.12a), and an object reading, as in example (4.12b). [17]

(4.12) a. tradició/tasca/institució/centre docent
tradition/task/institution/center teaching
'teaching tradition/task/institution/center'

b. planificació econòmica i docent
planning economical and teaching
'planning of economical aspects and teaching task'

c. equip docent
team teaching
'team of teachers'

In (4.12a), the meaning of the adjective can be paraphrased as "teaching", with an active event in its meaning. In examples (4.12b) and (4.12c), the task or the people involved in the teaching activity are focused on instead, which indicates an object-related sense. Note that in example (4.12b), the adjective is coordinated with an object-related adjective (*econòmic*). Although *docència* ('teaching activity') and *docent* ('teacher') are in turn event-denoting or event-related nouns, the event and object readings are distinguishable in many cases. Therefore, the adjective was classified as polysemous between event and object.

Polysemy between event- and object-related readings is the least frequent in the data: it only applies to six adjectives, namely, *digestiu*, *docent*, *comptable*, *cooperatiu*, *nutritiu*, and *vegetatiu*. In all these cases, the adjective bears a morphological relationship with a verb and/or a semantic relationship to an event. Because cases exist of basic-object, basic-event, and event-object polysemy, the way is opened to a three-way polysemy between basic, event- and object-related readings.

One such case could be *cooperatiu* ('cooperative'), which can be viewed as an attribute (basic reading, with a meaning similar to 'engaged'), an event-related meaning ('that cooperates'), and an object-related meaning ('related to cooperatives'). This is the only candidate to three-way polysemy we have found in our Gold Standard, so that it is an unfrequent phenomenon.

[17] Examples taken from the working corpus. Note that this is a case of synchronically not derived adjective with an event-related meaning, for the verb **docer* does not exist in Catalan. It was imported from Latin *docĕnte* (same meaning) directly into Catalan (Alcover and Moll, 2002).

We would like to compare the data obtained from the web experiment with the classification built by the three experts, so as to shed further light on the experiment and the sources of the disagreements. To facilitate comparison, a consensus classification was built using the participant's data. In the remaining of this section, we analyse the differences between the two classifications.

The simplest means to achieve a consensus classification is majority voting. We represent the semantic class of an adjective as the proportion of judgments provided by the participants corresponding to each of the six classes, and assign the adjective to the most voted class. The representation obtained for three of the adjectives in our Gold Standard is shown in Table 4.20.

Lemma	**Trans.**	**B**	**BE**	**BO**	**E**	**EO**	**O**
cranià	cranial	0	0	0	0	0	1
conservador	conservative	0.50	0.33	0.02	0.11	0.04	0
capaç	able	0.06	0.11	0.39	0.17	0.03	0.25

Table 4.20: *Examples for the representation of participants' classification.*

100% of the participants assigned *cranià* to the object class. For *conservador*, the judgments are more spread, but still half of the votes are concentrated in the basic class and a futher third in the basic-event class. Finally, for *capaç* the judgments are spread through all classes, with only a slight majority (39%) in the basic-object class.

The agreement scores between participants and experts are shown in Table 4.21. They are quite far from the desired 0.8 threshold, but they are much higher than the mean agreement between participants. The 0.55 K value is double as high as the 0.27 mean K value among participants, and wK reaches 0.65. The fact that individual biases are avoided in using a majority voting procedure yields a more stable classification, which corresponds better to the one provided by the experts. Note, however, that this does not add to the reliability of the Gold Standard. The reasons are that the participants' classification was obtained through a voting procedure and that the expert took into account participants' data in their classification.

p_o	K	wp_o	wK	op_o	oK
0.68	0.55	0.79	0.65	0.85	0.72

Table 4.21: *Agreement values: experts vs. participants.*

A useful tool to investigate the sources of the disagreements is the contingency table. Table 4.22 shows the contingency table obtained when the classifications built by the experts and the participants are compared.

In Table 4.22, the cells with highest values are bold faced. The highest values are found in the diagonal, but only for monosemous classes (B, E, and O), which indicates that there is a basic consensus on what the classes mean. However, high values are also found in two off-diagonal cells, namely, for lemmata which experts have tagged as basic and participants as object (B-O cell in Table 4.22), and for lemmata which experts have tagged as polysemous between basic and object and participants have tagged plainly as object (BO-O). Indeed, participants tend to assign many more lemmata to the object class than experts (note the difference in marginals: 73 for participants and 30 for experts). The rater bias test presented in Equation (4.6) (page (4.6)) confirms that the class distributions are significantly different (χ^2(1) = 33; p $< 10^-8$).

In the case of experts assigning basic and participants object (B-O), the following lemmata

		Participants						
		B	BE	BO	E	EO	O	*Total*
Experts	B	**79**	0	3	5	0	**20**	107
	BE	3	0	0	4	0	0	7
	BO	1	0	4	0	1	**17**	23
	E	2	1	1	**28**	1	4	37
	EO	0	0	0	2	2	2	6
	O	0	0	0	0	0	**30**	30
	Total	85	1	8	39	4	73	210

Table 4.22: *Contingency table: experts vs. participants.*

are involved: *calb* (‘bald’), *contingent* (‘contingent’), *desproporcionat* (‘disproportionate’), *intel·ligent* (‘intelligent’), *mal* (‘bad’), *morat* (‘purple’), *paradoxal* (‘paradoxical’), *perillós* (‘dangerous’), *pròsper* (‘prosperous’), *quadrat* (‘square’), *recíproc* (‘reciprocal’), *sant* (‘holy’), *semicircular* (‘semicircular’), *seriós* (‘serious’), *subterrani* (‘underground’), *titular* (‘titular’), *triangular* (‘triangular’), *viciós* (‘viciós’), *vigorós* (‘vigorous’), *viril* (‘virile’).

For many of these cases, there exists a deadjectival noun corresponding to the attribute denoted by the adjective: *calbície* (‘baldness’) for *calb*, *intel·ligència* (‘intelligence’) for *intel·ligent*, *reciprocitat* (‘reciprocity’) for *recíproc*, *santedat* (‘holiness’) for *sant*. These nouns denote attributes and not objects, and the “related to” pattern cannot be properly applied to the adjectives to describe their meaning. The adjective *calb*, for instance, does not **mean** “related to baldness”, which is what the use of the object pattern implies, although the meaning of *calb* is indeed related to the meaning of *calbície*.

The behaviour of the participants suggests that attribute-denoting deadjectival nouns are particularly salient for these adjectives. Recall that in WordNet (see Section 3.5.1), links indicating relationships between adjectives and derived nouns are explicitly coded. Also, from their behaviour it seems that a suitable synonym or antonym (indicative of the basic class) is not so salient as the derived noun.

In the case of more prototypical basic adjectives, the reverse is true. For instance, for *ample* (‘wide’) the deadjectival nouns *amplada, amplària* and *amplitud* (‘wideness’) exist, and they have been provided by 18 out of the 58 judges classifying this adjective. However, the antonym *estret* (‘narrow’) is so readily available that an overwhelming majority of responses, 49 out of 58, include it (many participants however provided multiple responses). This piece of evidence supports the claim made in Chapter 3 that the synonymy/antonymy lexical relationship accounts only for the most prototypical basic adjectives.

Because deciding whether a noun refers to an attribute or an object is again a subjective decision, the filtering procedure explained in Section 4.2.4 did not filter these cases out. It is clear, however, that in many cases the usage of this pattern did not correspond to its intended use, and that the design of the experiment should be worked upon to avoid this confusion.

For the case of experts assigning basic-object and participants only object (BO-O cell in Table 4.22), the following lemmata are involved: *amorós* (‘affectionate|of love’), *anarquista* (‘anarchist(ic)’), *capitalista* (‘capitalist(ic)’), *catalanista* (‘that supports Catalan autonomy’), *comunista* (‘communist’), *eròtic* (‘erotic’), *familiar* (‘familiar|of family’), *humà* (‘human’), *intuïtiu* (‘intuitive|of intuition’), *local* (‘local’), *poètic* (‘poetic(al)’), *professional* (‘professional’), *sen-*

sitiu ('sentient|sensitive'), *socialista* ('socialist(ic)'), *turístic* ('tourist|touristy'), *unitari* ('unitary|of unit'), *utilitari* ('utilitarian|of utility').

Many of these adjectives are of the ideology type discussed above, which have been consistently tagged as basic-object by the experts because they seem to be ambiguous between an attribute reading (mostly when applied to humans) and a relationship with an object reading (the abstract object corresponding to the underlying ideology). The expert judges considered the representation in terms of basic-object to be the best representation for this ambiguity in our setting. However, it is not the optimal treatment and it needed an explicit convention. Most participants simply included the relationship to the ideology.

The remaining cases mostly correspond to true polysemy, mainly object-related adjectives that have acquired a basic reading, as discussed in Section 3.6. Example (3.39b) (page 54), which exemplifies the two senses of *familiar,* is repeated here for clarity purposes as (4.13).

(4.13) reunió familiar / cara familiar
meeting familiar / face familiar

'family meeting / familiar face'

The translations given above clarify the polysemy captured by the polysemous class assignments for many adjectives. Because these adjectives are denominal, participants tended to provide only the object reading and gloss over the basic reading.

For *nocturn* ('nocturnal|of night') and *diürn* ('diurnal|of day'), however, the majority assignment of participants coincides with that of experts, namely, ambiguous between basic and object. The relationship to objects *night* and *day,* as well as the antonymy relationship between *nocturn* and *diürn*, were strong enough to make the majority of votes assign BO, even if the lemmata were in different test sets.

It was intended for the participants to provide multiple assignments in case of polysemy, as with the *nocturn/diürn* case. However, in general, they provided multiple responses in difficult or ambiguous cases instead. In the cases in which participants consistently provided ambiguous readings (which were very few), this did not indicate polysemy. One such case is *capaç*.

The judgment distribution across classes of *capaç* is given in Table 4.20 above, and is very spread. The most voted class (39%) is the basic-object class. However, the most frequent answers (*incapaç* 'unable' for the basic pattern and *capacitat* 'ability' for the object pattern) do not point to different senses. They rather suggest that the judges could not make their minds up with respect to the class of the adjective.

In fact, although individually participants could provide many multiple answers (depending on personal taste or understanding of the task), as a collective they assign almost exclusively monosemous classes, which indicates wide disagreement in the use of polysemous classes. The cases where experts provide a polysemous class and participants a monosemous class are underlined in Table 4.22, and they constitute the third main source of disagreement.

Out of the 7 cases tagged as basic-event by experts (second row in Table 4.22), three are assigned to basic and four to event by the participants. Similarly, of the 23 BO cases according to experts, one is disambiguated as basic, four remain BO, and 17 are assigned to object only, as we have just discussed. Also, of the 6 lemmata classified as EO by experts, two are disambiguated as event, two as object, and other two remain EO according to participants.

Finally, note that except for the two off-diagonal bold faced cells (B-O and BO-O), most of the cases of disagreement involve the event class. Out of the 67 cases where experts and participants disagree with respect to the semantic class of the adjectives, 28 involve the event class (that is, involve classes BE, E, and EO). Of the remaining 41 cases, 37 correspond to the B-O and BO-O cases explained above. We have argued that B-O and BO-O disagreements are due to experimental design problems (which caused confusion with the use of the object pattern) and to the nonconsistent use of multiple responses to encode polysemy judgments. However, the categories basic and object seem to be well defined apart from this misunderstanding.

In contrast, disagreements involving the event class cause small numbers to appear all over Table 4.22, which can be viewed as random disagreements indicating confusion with respect to the definition of the event class. This hypothesis, together with hypotheses regarding the other sources of disagreement discussed in this section, will be tested in the next section.

4.5.2 Using entropy to measure the difficulty of an adjective

The representation of the participants' judgments about Gold Standard presented in the Section 4.5.1 allows for a further kind of inter-item analysis. Because we have a frequency distribution for each adjective, we can assess the degree of coincidence of judges according to the amount of probability mass assigned to each class. Intuitively, if all the mass is concentrated in one class, there is total coincidence of judgments; if it is evenly spread, there is no consensus.

A simple way of formalising this notion is to use the Information Theory measure of entropy introduced by Shannon (1948). [18] Entropy measures the average uncertainty in a random variable. If X is a discrete random variable with probability distribution $p(x) = Pr(X = x)$ (x being each of the possible outcomes of the variable), its entropy is computed as in Equation (4.13). The equation specifies log base 2, although entropy can be computed using other bases, because it is usually measured in bits, and we adhere here to this convention. If the outcome of the variable is totally predictable, the uncertainty (and thus the entropy) is 0, and as the unpredictability increases, entropy also increases, with an upper bound determined by the number of possible outcomes of the random variable.

$$H(X) = -\sum p(x)\ \log_2 p(x) \tag{4.13}$$

In our case, the random variable is the class of the adjective, and predictability amounts to coincidence among judges. Table 4.23 shows that the measure intuitively corresponds to what it aims at measuring: for *cranià*, with total coincidence, entropy is 0; for *conservador*, with a half of the probability mass in a class (B) and one third in another class (BE), entropy increases to 1.17. And finally, for *capaç*, with very spread judgments, it increases to 1.52. The upper bound for entropy in our case is 2.58, for the case when all classes have an equal probability, 1/6[19]. However, the maximum entropy reached for an adjective in our data was 1.74 (adjectiu *orientat*, 'oriented').

[18] An alternative would be the intra-item score obtained from multi-judge agreement measures such as α or β. These measures amount to mean intra-item agreement scores, as discussed in Artstein and Poesio (2005b). Intra-item agreement can be measured in different ways depending on the assumptions made about the underlying distribution of categories. We expect the information provided by entropy and by intra-item agreement to be roughly the same.

[19] $p(x) = 1/6$; $H(class) = -6(1/6)\log_2(1/6) = -\log_2(1/6) = 2.58$.

The maximum value could be used as a baseline to compare entropy results to. However, the assumption of homogeneous distribution is too strong: different classes clearly appear in varying frequencies in the web experiment data. Therefore, the maximum value is not an adequate baseline. An alternative baseline is to estimate the entropy of the class distribution, namely, the distribution obtained by averaging all individual class distributions. The two last rows in Table 4.23 show the data for the baseline and the maximum entropy estimates.

Lemma	**Trans.**	**B**	**BE**	**BO**	**E**	**EO**	**O**	**Entropy**
cranià	cranial	0	0	0	0	0	1	0
conservador	conservative	0.5	0.33	0.02	0.11	0.04	0	1.17
capaç	able	0.06	0.11	0.39	0.17	0.03	0.25	1.52
baseline	-	0.30	0.06	0.13	0.14	0.06	0.30	2.32
maximum	-	1/6	1/6	1/6	1/6	1/6	1/6	2.58

Table 4.23: *Entropy values from participants' classification.*

Entropy seems to successfully capture variations in agreement. The question arises of whether it could be used plainly as an agreement measure. The fact that all it needs is a probability distribution makes it easy to merge data obtained from different judges. In our case, the 7 test sets can be considered as a single set, and thus obtain a single agreement measure for the task in a straightforward way. However, the fact that the subset of data from the same set of judges is not independent of one another raises difficulties to estimate confidence intervals. To use entropy as an agreement measure, the items should be randomised and not grouped into distinct test sets.

Another disadvantage is that it does not have a uniform upper bound, but it depends on the number of classes in the study, so that it is difficult to compare results across studies. Finally, one practical shortcoming is that, to obtain reliable probability estimates, a quite large number of judges must be used, which is probably the reason why it is not often used as an agreement measure. Nevertheless, it has recently been used as an agreement or consensus measure in at least one educational article (Tastle and Russell, 2003).

As can be seen in Figure 4.1, which depicts the distribution of adjective entropy values, the picture that emerges as to the agreement achieved among human judges is similar to the analysis made in Section 4.4. If the levels of agreement were optimal, the histogram would be skewed right, and most values would pile up around 0. We see that it is skewed left, and that most entropy values are central values (mean is 1.07, standard deviation 0.37), if we consider that possible values for our task range from 0 to 2.58 and that the baseline is 2.32. All obtained values are well below the baseline, so that there is evidence for some degree of agreement, but it represents moderate agreement.

Whether or not it makes sense as an agreement measure, entropy can clearly be used for inter-item analysis, namely, to analyse whether there are types of objects that consistently have a higher or lower entropy value, which can be taken to indicate a higher or lower difficulty of the type of adjective. Differing judgments result in a more spread distribution of values across classes. A more spread distribution results in a higher entropy value. Therefore, a higher entropy value indicates a higher difficulty or confusion with respect to a given adjective.

Several explanations for differences in entropy values can be envisaged. We next assess the sources of disagreement that were discussed in Section 4.5.1. Let us first discuss polysemy and disagreements with experts.

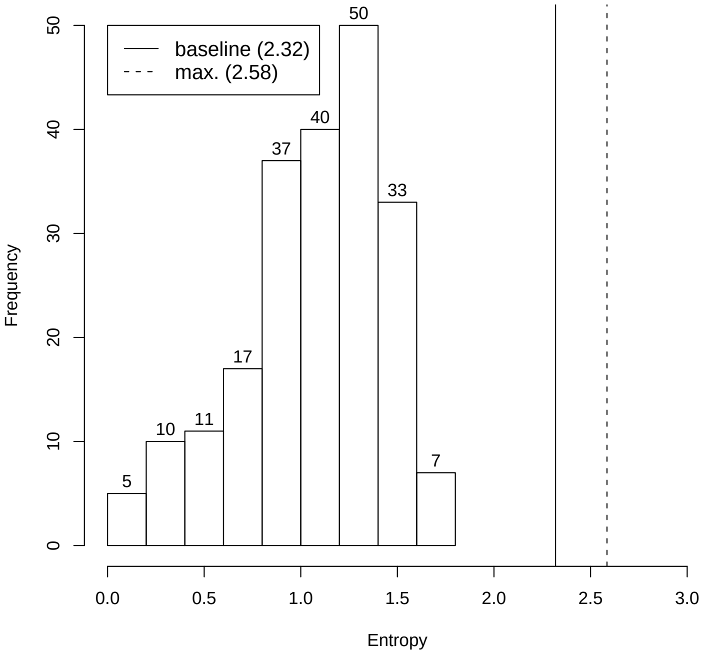


Figure 4.1: *Distribution of entropy values for adjectives.*

Adjectives classified by the experts as polysemous should yield less compact judgments than monosemous adjectives, and therefore have a higher entropy. This could have two sources. First, the participants could be unsure as to which class or classes the adjective belongs to. Second, some participants could code the class corresponding to only one of the senses (in either of the relevant monosemous classes, the one that is most salient for them), and some others could code the polysemous class corresponding to the two senses. For instance, for a BE adjective, some participants could code it as B, others as BE, and still others as E. Because they are considered as separate classes, this would yield higher entropy values for these adjectives than for monosemous adjectives.

Similarly, cases where participants and experts disagree can be expected to be more controversial than cases where there is agreement. The same reasoning applies: adjectives that are difficult or do not fit in the classification should yield more spread distributions of values across classes in the participants' classification. Also, a somewhat arbitrary decision is likely to be made in both the participants' and the experts' classifications, which is likely to cause mismatches betweeen them. Therefore, we expect adjectives for which experts and participants disagree to exhibit higher entropy values.

Figure 4.2 shows that both of the predictions outlined are met.

Polysemous adjectives have higher entropy (mean = 1.2, standard deviation = 0.29) than monosemous adjectives (M = 1.05, SD = 0.38). The difference is significant (t(62.3) = -2.6, p = 0.01, two-tailed). [20]

Adjectives for which participants and experts disagree with respect to their semantic class also

[20] Equality of variance is not assumed.

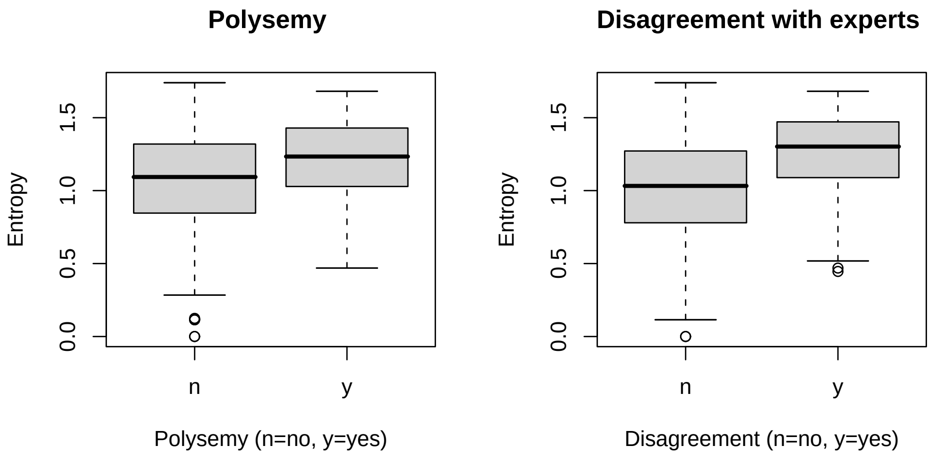


Figure 4.2: *Explaining differences in entropy values I.*

exhibit higher entropy (M = 1.25, SD = 0.30) than those for which there is agreement (M = 0.99, SD = 0.38). The difference is again significant (t(160.2) = -5.28, p $< 10^{-}6$, two-tailed). Note that the differences in entropy values are higher for the second explanation than for the first one, which means that disagreements between experts and participants predict difficulty to a larger extent than polysemy.

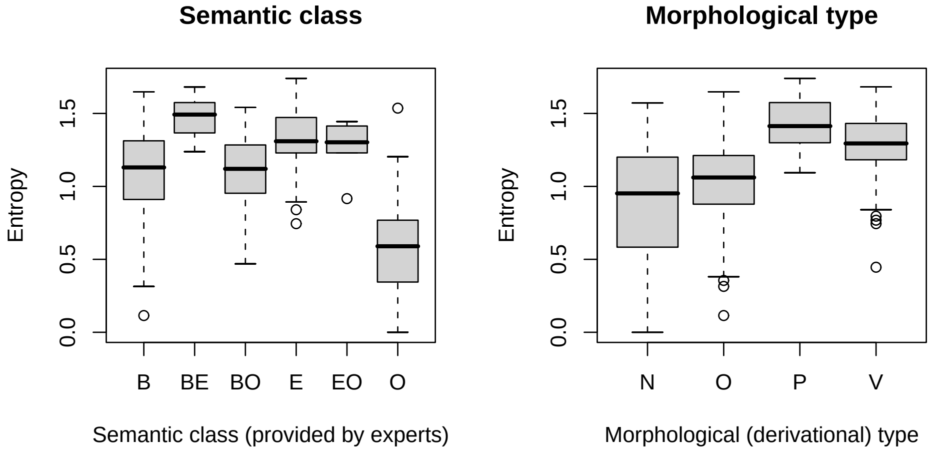


Figure 4.3: *Explaining differences in entropy values II.*

At the end of Section 4.5.1, we have made the hypothesis that, despite the fact that disagreements seem to be concentrated in the basic vs. object distinction, this responds to a different understanding of the object class definition, not to the inherent difficulty of the adjectives. Instead, there were numerous small disagreements concerning the event class, which, we argued, could point to a confusion with respect to the definition of the class.

Entropy provides us with a means to test this explanation. If it is along the right track, adjectives classified as event by the experts should have higher entropy values, because the participants should be more unsure as to where to place them, resulting in more spread distributions of values across classes. The first graphic in Figure 4.3 shows that event-related adjectives (classes

BE, E, EO) are clearly more controversial than the rest, supporting the explanation. It also shows that object-related adjectives (class O) are the least problematic ones. Both pieces of evidence are consistent with the data regarding class-specific agreement (Table 4.18, page 90) discussed in Section 4.4. One-way ANOVA confirms that mean entropy values are different depending on the class (F(5, 29.3) = 23.1, p $< 10^{-}8$). [21]

Our hypothesis is that event adjectives are problematic due to two main factors. One is the fact that the semantics contributed by the morphology is much more diverse than that of object adjectives. In our manually annotated database of adjectives, there are only 8 different suffixes for deverbal adjectives, while there are 22 for denominal adjectives. However, object adjectives show a much more compact semantics than event adjectives, as shown by the fact that we defined 3 patterns to account for the semantics of event adjectives, and only one for object (and basic) adjectives.

The second factor is that the semantics contributed by the deriving verb also results in high variability, mainly due to the *Aktionsart* of the verb. Stative verbs produce more “basic-like” event adjectives. For instance, *abundant* was classified as event by the experts due to its relationship with the verb *abundar*. It was classified as basic by the participants due to it being antonymous to *escàs* (‘sparse’), mirroring the fact that it denotes an attribute, like a basic adjective. Adjectives derived from process-denoting verbs have a more distinct semantics.

Because semantic class and morphological type of an adjective are related, as argued in Chapter 3, we expect the differences to map to the morphological level. The second graphic in Figure 4.3 shows that participial and deverbal adjectives, those that correlate with the event class, have higher entropy values than the rest, so that they are more controversial. [22] The results of an ANOVA test again confirm this analysis (F(3, 68.4) = 27.1, p $< 10^{-}10$). Not derived adjectives are somewhat more controversial than denominal adjectives (confirmed by a t-test: t(131.6), p = 0.02). This piece of evidence is in accordance with basic adjectives being more controversial than object adjectives, as shown in the graphic *Semantic class* in Figure 4.3. Note, however, that the difference is clearer at the semantic level (for basic against object, t(42.3) = 7.03, p < 10^{-07}).

Table 4.24 shows the mean (M) and standard deviation (SD) values for the data corresponding to Figures 4.2 and 4.3. The means for the disagreement and not disagreement variable are 0.99 and 1.25, respectively. If disagreements within the B-BO-O categories (as opposed to BE-E-EO classes) are excluded from the analysis, the entropy mean for adjectives with disagreement rises to 1.41. Correspondingly, the p value of the significance test decreases from p $< 10^{-}6$ to $< 10^{-}11$ (t(67.5) = -8.59). This adds support to the argument that the event class is a key piece of evidence in explaining coding difficulties, suggesting that it is the least clearly defined.

4.6 Summary

Web experiments such as the one explained in this chapter are a promising way to gather linguistic data. They provide clues for research on linguistics that would be very hard, if not

[21] Homogeneity of variance is not assumed for the ANOVAs performed in this Chapter.

[22] Legend for graphic *Morphological type* in Figure 4.3:

N	denominal
O	not derived
P	participial
V	deverbal

Disagreement	**n**	**y**				
M	0.99	1.25				
SD	0.38	0.30				
Polysemous	**n**	**y**				
M	1.05	1.20				
SD	0.38	0.29				
Semantic class	**B**	**BE**	**BO**	**E**	**EO**	**O**
M	1.09	1.47	1.09	1.31	1.27	0.58
SD	0.31	0.17	0.29	0.23	0.19	0.36
Morph. type	**N**	**O**	**P**	**V**		
M	0.89	1.03	1.44	1.26		
SD	0.41	0.32	0.19	0.25		

Table 4.24: *Mean and standard deviation according to adjective type.*

impossible, to obtain from introspection or corpus data. For computational linguistics, they provide the possibility to evaluate agreement using many more judges than is usually done. This opens the way to more robust Gold Standard resources, as well as to an improvement of the definition of both the categories involved and the tagging tasks. However, these experiments are extremely difficult to design, particularly if they are addressed to naive subjects, as opposed to a few chosen experts in a field. This is a necessary requirement if a large number of judgments are to be obtained. Web experiments also open the way to different methodologies for the estimation of agreement than are traditional for K and related measures, as our discussion of the estimation of population mean has shown.

The interrater agreement for our task, with the current experimental design, is very low, much too low for academic standards. This is particularly evident for the strictest evaluation mode (*full* agreement), strictly including polysemy judgments. We expected judges to assign multiple categories to lemmata in case of polysemy. The analysis of the agreement data has shown that they do not do so, but that they rather code either ambiguous or difficult cases using multiple categories. Reliable acquisition of polysemy judgments for our task remains an unresolved challenge, and has repeatedly proven difficult to achieve in related research.

The analysis of the patterns of disagreement has revealed confusion in using the object pattern. Attribute-denoting nouns are very salient for many basic adjectives, and suitable synonyms or antonyms (the cue to identify basic adjectives) are not always available. Despite this confusion, adjectives belonging to the basic and object classes seem not to be highly problematic, while adjectives for which the event class is involved are significantly more controversial than the rest. We could measure this aspect through the use of entropy as a measure of lemma-specific agreement.

The low level of agreement for our experiment, thus, has to do both with the design of the experiment and with difficulties in the classification. As for the latter aspect, the analysis suggests that the event class is the less clearly defined of the three classes, a result that is supported by the machine learning experiments presented in Chapter 6.

As for the design of the experiment, the main problem is that, although it is directed at naive subjects, it asks for metalinguistic judgments. Building dictionary definitions is not an intuitive task. This could explain the high dropout rate for the experiment. A task demanding linguistic intuitions should be designed instead, so that a proper psycholinguistic experiment could be carried out. How to best define such a task remains an open question.

Due to the low agreement among our judges, and to the fact that the usage of the patterns did not always correspond to their intended use, we will use the expert classification as a Gold Standard for the supervised experiments explained in Chapter 6. Before turning to that, initial unsupervised experiments designed to assess and refine the classification will be explained in the next chapter.

Finally, the following table summarises the characteristics of the 3 Gold Standards built during the PhD, which we will label A, B, and C, and describes their role in the machine learning experiments of Chapters 5 and 6.[23] The full list of adjectives for the three Gold Standards and the corresponding data is in Appendix C.

	A	**B**	**C**
#judges	4	3	322 + 3 experts
#lemmata	101	80	210
mean K	0.58	0.74	0.38
main class	intensional, qualitative, relational	basic, event, object	basic, event, object
polys. classes	intens.-qual., qual.-rel.	*none*	bas.-ev., bas.-obj., ev.-obj.
used in	Exp. A (Section 5.1)	Exp. B (Section 5.2)	Exp. C (Chapter 6)

Table 4.25: *Gold Standards built for machine learning experiments.*

Part of the material presented in this chapter has been submitted for publication in a journal.

[23] Legend:

GS:	Gold Standard
main cl.:	main classification
polys.:	polysemous classes
Exp.:	experiment

Chapter 5
Experiments A and B: Assessing the classification

> . . . we view corpora, especially if annotated with currently available tools, as repositories of implicit grammars, which can be exploited in automatic [classification] tasks.
>
> Merlo and Stevenson (2001, p. 399)

This chapter reports on a set of unsupervised experiments performed with three main goals: first, to provide feedback to the proposed classification. As has been discussed in Chapter 3, it is not clear from a theoretical point of view how to best classify adjectives according to their lexical semantics. It makes sense to look for large-scale empirical evidence for this problem, something that the methods developed within Lexical Acquisition offer.

A second, related goal is to see how polysemy fits in the overall picture: how polysemous adjectives behave with respect to monosemous adjectives, and how the clustering algorithm behaves with respect to polysemous adjectives.

The third main goal is to test different representations for the relevant distributional properties of adjectives that correlate with their semantic class, so as to test their relative performance in using them for machine learning experiments. This should also serve to provide a first sketch of the classes in terms of the different feature representations chosen.

To meet these goals, unsupervised techniques, and in particular clustering, are a natural choice. Supervised techniques use training data, already labeled, to learn a model of the different classes, and use this model to tag unseen data. With clustering, no labeling of training data is necessary. Objects are grouped together according to their feature value distribution, not to a predefined classification. Potentially, a better insight into the structures present in the data can be achieved. Unsupervised techniques can thus be viewed as less biased than supervised techniques, although they are obviously biased through the choice of feature representation. For instance, a grouping of people in terms of height and weight will result in a different classification than a grouping based on hair and eye colour. The specific clustering algorithm also influences the resulting structure: some algorithms favour globular clusters, others elongated ones, etc.

One of the consequences of using an unsupervised methodology is that an arbitrarily large number of objects can be used for classification, because they do not need to be labeled. In all the experiments reported in this chapter, the whole set of adjectives meeting the criteria are clustered; however, the results are analysed using limited sets of labeled data. We compare the results with classifications established by human judges according to semantic characteristics.

Adjectives have a much more limited distribution than verbs or nouns, and do not usually present long-distance dependencies. They are basically restricted to modification within NP and predicative constructions such as copular sentences (see Chapter 3). Therefore, we expect that distributional features within a small window will provide enough evidence for our task. This would facilitate acquisition experiments for languages with no widely available deep-processing resources.

The features used mainly model the syntactic behaviour of adjectives. Our approach exploits the syntax-semantics interface as is usual in Lexical Acquisition: assuming that semantic similarities drive to syntactic similarities, it is possible to go the opposite way and induce semantic similarities from syntactic similarities. We also begin to explore the morphology-semantics interface, although a more complete exploration is offered in Chapter 6.

Section 5.1 describes initial clustering experiments, which lead to a revision of the initially proposed classification. Section 5.2 explains further unsupervised experiments with the final semantic classification, which confirm the plausibility of the classification and at the same time uncover new problems concerning the semantic classes.

5.1 Experiment A: refining the classification

5.1.1 Classification and Gold Standard

In this experiment, the classification distinguishes between qualitative, relational, and intensional adjectives (see Chapter 3 for an explanation of these terms). Polysemous adjectives are assigned to "polysemous" classes qualitative/relational, and qualitative/intensional. No cases of intensional/relational were found.

To analyse results, we use the first Gold Standard established within the PhD (Gold Standard A in Table 4.25, page 4.25). As explained in Section 4.1, this Gold Standard consists of 101 adjective lemmata, chosen among the 3,521 lemmata with more than 10 occurences in the study corpus. [1] 99 lemmata were classified by 4 human judges, with a mean K value of 0.58, and two intensional adjectives were subsequently added, because the class was not represented in the Gold Standard. The 4 different classifications were merged into a single Gold Standard by the author of this thesis so as to have a unique classification to compare the machine learning results to.

5.1.2 Features

Adjective lemmata are modeled using two kinds of shallow cues. First, we defined textual correlates for some of the theoretically relevant properties of each class reviewed in Chapter 3. In addition, because of the exploratory nature of this experiment, we built a relatively neutral model, so as to blindly model the syntactic distribution of adjectives. The second representation takes into account the n-gram distribution of adjectives, defined in terms of the POS of the surrounding words. We now describe each feature type in more detail.

[1] The corpus corresponds to roughly half (8 million words) of the CTILC fragment used for subsequent experiments, because only that fragment was available at the time of the experiments.

5.1.2.1 Semantic features

The features based on theoretical considerations will be referred to as *semantic features* in what follows. Although they are defined in terms of shallow cues, they are textual correlates of mainly semantic properties. For instance, the presence of a particular type of adverb to the left of an adjective is an indication of its being gradable, a semantic property. However, the different linguistic levels of description are closely tangled: for instance, a predicative syntactic function is an indication that the adjective can semantically function as a predicate. Whether the feature is labeled as semantic or syntactic depends on the rest of the features in the particular linguistic level.

The list of semantic features extracted, selected according to the considerations in Chapter 3 and defined as shallow cues of linguistic properties, were the following.

Gradable Adjective preceded by a degree modifier[2], occuring with a degree suffix (such as *-et*, *-ó*, *-íssim*), or coordinated with an adjective preceded by degree modifier or exhibiting a degree suffix.

Comparable Adjective preceded by degree modifiers *més* ('more') or *menys* ('less'), or preceded by *tan* and followed by *com* (comparative construction). In Catalan there is no comparative inflection.

Predicate Two features are distinguished: adjective acting as a predicate in a copular sentence (feature *copular*) and in other constructions (feature *pred*). This distinction is made in the functional tagset of CatCG (see Table 3.1 in page 20).

Not restrictive As has been explained in Section 3.6.1.1, non restrictive adjectives usually precede the head noun. The pre-nominal modification function tag assigned by CatCG (Table 3.1) is the indicator chosen for this feature.

Adjacent As has been explained in Section 3.6.1.3, the ordering of adjectives within an NP can be a useful cue to their semantic class, for relational adjectives tend to occur closer to the head noun than qualitative ones. Ocurrences of adjectives following a noun and preceding an adjective were taken as indicative of this property, with the following constraints: the noun and the two adjectives should agree, the second adjective could be preceded or not by degree modifiers, and it had to bear the same function tag as the first adjective.

Some of these features are clearly related: for instance, *gradable* and *comparable*, or *copular* and *pred*. However, there could be differences between their distributions across different classes, so that they were kept separate for exploration.

Table 5.1 summarises the features chosen, together with their mean and standard deviation values. For each adjective, the feature values are encoded as proportions of occurences in each

[2] Defined as a list of adverbs: *bastant, ben, força, gaire, gens, gota, massa, mica, mig, molt, poc, prou, tan, altament, completament, considerablement, enormement, extremadament, extremament, immensament, infinitament, lleugerament, mínimament, mitjanament, moderadament, parcialment, relativament, sensiblement, summament, terriblement, totalment, tremendament.*

of the defined contexts. Note that the mean values are very low, which will be taken into account in the analysis. The standard deviations, in contrast, are quite large relative to mean values (about double as high as the mean), due to the Zipfean distribution of features.

Feature	Textual correlate	Mean	SD
gradable	degree adverbs, degree suffixation	0.04	0.08
comparable	comparative constructions	0.03	0.07
copular	copular predicate syntactic tag	0.06	0.10
pred	predicate syntactic tag	0.03	0.06
not restrictive	pre-nominal modifier syntactic tag	0.04	0.08
adjacent	first adjective in a series of two or more	0.03	0.05

Table 5.1: *Semantic features.*

From the discussion in Chapter 3, the following predictions with respect to the semantic features can be made:

- In comparison with the other classes, qualitative adjectives should have higher values in features *gradable*, *comparable*, *copular*, *pred*, middle values in feature *not restrictive* (lower than intensional adjectives and higher than relational adjectives), and low values for feature *adjacent*.

- Relational adjectives should have the opposite distribution, with very low values for all features except for *adjacent*.

- Intensional adjectives are expected to exhibit very low values for all features except for *not restrictive*, for which a very high value is expected.

- With respect to polysemous adjectives, it can be foreseen that their feature values will be in between those of "canonical" classes. For instance, an adjective that is polysemous between a qualitative and a relational reading (such as *familiar*) should have values for feature *gradable* that are higher than for a monosemous relational adjective but lower than a typical qualitative adjective.

Figure 5.1 shows that the predictions just outlined are met to a large extent. [3]

The differences in value distribution are mostly not sharp (most of the ranges in the boxes overlap). This affects mainly polysemous classes: although they show the tendency just predicted of exhibiting values that are in between those of the main classes, they are not differentiated. The clustering results will be affected by this distribution, as will be discussed in Section 5.1.5.

[3] Legend for class labels:

I:	intensional
IQ:	polysemous between intensional and qualitative
Q:	qualitative
QR:	polysemous between qualitative and relational
R:	relational

Note that the scale in Figure 5.1 does not range from 0 to 1; this is because the data are standardised, as will be explained in Section 5.1.3.

However, one-way ANOVA tests on each of the features (factor: classes), excluding items in the I and IQ classes because not enough observations are available,[4] yield significant results, with p lower than 0.05 (*pred*), 0.01 (*comparable, not restrictive, adjacent*), and 0.001 (*gradable, copular*). Significant results are due mainly to differences between qualitative and relational classes. The full data for the ANOVA tests is shown in Table 5.2.

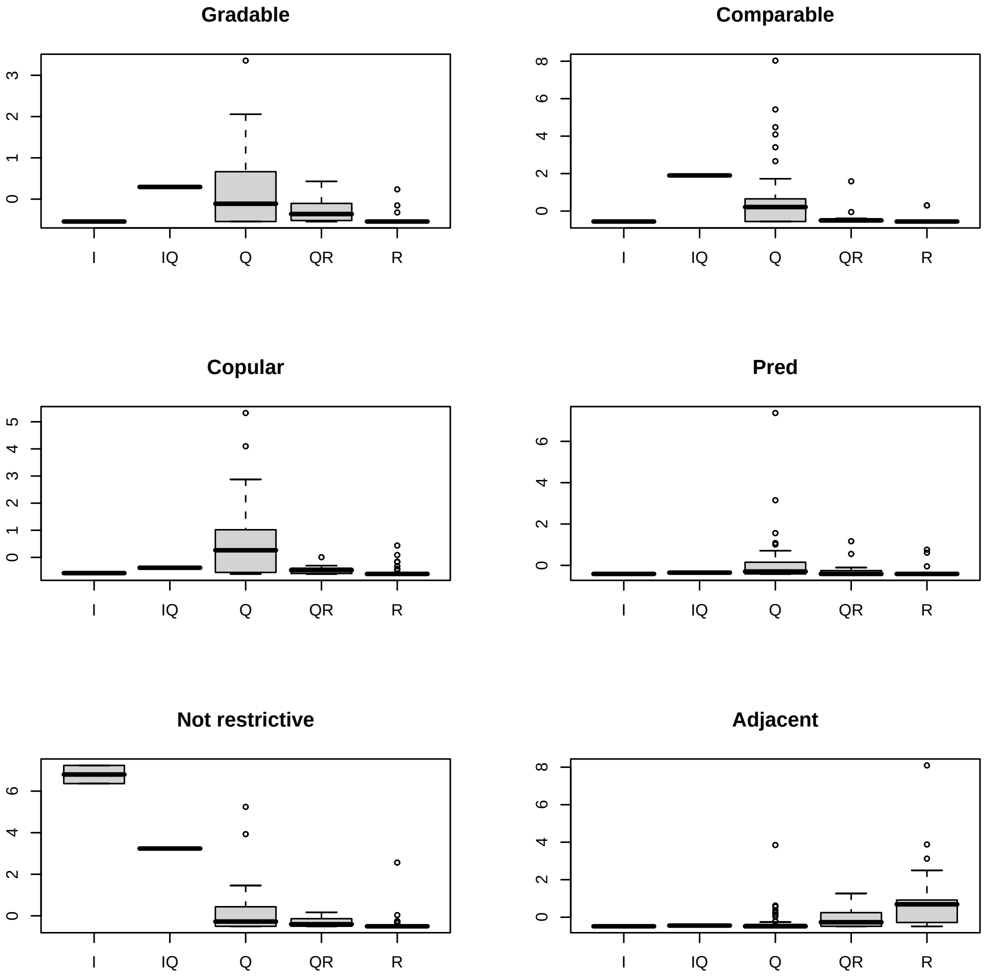


Figure 5.1: *Feature value distribution for semantic features (classes).*

5.1.2.2 Distributional features

For distributional features, the POS of two words at each side of the adjective (5-word window) are recorded as separate features. For instance, for an occurence of *fix* 'fixed' as in (5.1a), the representation would be as in (5.1b). In the example, the target adjective is in bold face, and the relevant word window is in italics. Negative numbers indicate positions to the left, positive ones positions to the right. The representation in (5.1b) corresponds to the POS of *molt*, *menys*, , (comma), and *o*.

[4] Only two items (*mer* 'mere', *presumpte* 'alleged') are in the intensional class, and one (*antic*, 'old/former') in the intensional/qualitative polysemous class.

	Q	**QR**	**R**	**df**	**F**	**p**
Gradable	0.16 ±0.81	-0.27 ±0.31	-0.50 ±0.15	2,25.1	17.9	$< 10^{-4}$
Comparable	0.58 ±1.71	-0.28 ±0.64	-0.51 ±0.20	2,23.6	10.6	0.0005
Copular	0.54 ±1.27	-0.45 ±0.19	-0.52 ±0.23	2,38.9	16.9	$< 10^{-5}$
Pred	0.13 ±1.21	-0.15 ±0.53	-0.34 ±0.26	2,25.5	3.9	0.03
Not restrictive	0.18 ±1.07	-0.29 ±0.25	-0.38 ±0.52	2,52.0	5.2	0.008
Adjacent	-0.28 ±0.64	-0.06 ±0.58	0.86 ±1.66	2,28.9	7.6	0.002

Table 5.2: *Semantic features: differences across classes.*

(5.1) a. Els instints domèstics, per dir-ho així, són del cert *molt menys* **fixos**, *o*
The instincts domestic, to say-it so, are of-the true very less fixed, or
invariables, que els instints naturals
invariable, than the instincts natural

‘Domestic instincts, to put it this way, are really much less fixed, or invariable, than natural instincts.’

b. *-2 adverb, -1 adverb, +1 punctuation, +2 conjunction*

Our tagset distinguishes between nine POS: verb, noun, adjective, adverb, preposition, determiner, pronoun, conjunction, and punctuation. Because the POS of 4 different positions (two to the left, two to the right of the target adjective) are separately encoded, we end up with 36 different distributional features. The 10 features with the highest mean value are listed in Table 5.3.

Feature	**Mean**	**SD**
-1 noun	0.52	0.25
+1 punctuation	0.42	0.15
-2 determiner	0.39	0.20
+2 determiner	0.24	0.13
+1 preposition	0.21	0.15
-2 preposition	0.13	0.09
-1 adverb	0.10	0.11
-1 verb	0.08	0.11
-1 determiner	0.06	0.10
+1 noun	0.06	0.10

Table 5.3: *Distributional features.*

These features can be viewed as a very simple correlate of the syntactic behaviour of adjectives. For instance, from Table 5.3 it can be deduced that the default position of the adjective in Catalan is the postnominal one, because on average adjectives occur immediately after a noun in more than half their total occurences (52%). The 10 features in Table 5.3 have much higher values than the semantic features in Table 5.1. This is due to the fact that contexts corresponding to semantic features are not frequent, while all examples have a distribution in terms of the POS of their surrounding words.

5.1.3 Clustering parameters

Clustering (see Kaufman and Rousseeuw (1990) and Everitt, Landau and Leese (2001) for comprehensive introductions to this technique) is an Exploratory Data Analysis technique that forms groups (clusters) of homogeneous objects represented as a vector space. Each object that is to be grouped (in our case, each adjective lemma) is represented as a set of features with their associated values.

For instance, Table 5.4 contains the vectors for adjectives *malalt* ('ill'), *avergonyit* ('ashamed'), and *freudià* ('Freudian') represented with the semantic features. Each column (dimension) of the matrix corresponds to a feature, and the algorithm acts on each row (vector) of the matrix. In this case, the first dimension corresponds to feature *gradable*, and each of the rest to the remaining semantic features, in the same order as in Table 5.1 and Figure 5.1.

0.23	-0.47	2.65	-0.33	-0.45	-0.49
0.61	-0.56	-0.61	-0.41	-0.50	-0.49
-0.54	-0.56	-0.61	0.76	-0.50	0.85

Table 5.4: *Representation for adjectives* malalt, avergonyit, freudià.

The goal of clustering algorithms is to group similar objects together, and put dissimilar objects into separate groups. The similarity between objects is measured through the comparison of their feature values. There are many different measures and criteria to measure similarity, but they are all based on the notion of vector distance. [5]

Although there are dozens of clustering algorithms, the main techniques can be described with few parameters. One of the main parameters is whether the resulting clustering structure is hierarchical or flat. In hierarchical algorithms, clustering proceeds progressively, so that the decisions of the algorithm can be viewed as a tree (called a *dendogram*). In direct algorithms, which yield flat structures, the clusters are made once and for all.

Another relevant parameter for an algorithm is its agglomerative or partitional nature. In agglomerative algorithms, n clusters are initially built (where n is the total number of objects, so that each object constitutes a cluster), and these minimal clusters are grouped according to a similarity metric into the k desired clusters. Partitional algorithms work the other way round: they begin with a single cluster with all objects, and divide it into k clusters.

For the experiment, we tested several implementations of clustering algorithms provided in the CLUTO toolkit (see Section 2.2.2): 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. The overall picture of the structure present in the data was quite robust across different parametrisations. For clarity reasons, we will restrict the discussion to one parametrisation, corresponding to the k-means clustering algorithm. This is a classical algorithm, conceptually simple and computationally efficient, which has been used in related work, such as the induction of German semantic verb classes (Schulte im Walde, 2006) or the syntactic classification of verbs in Catalan (Mayol et al., 2005).

K-means is a flat, partitional algorithm which works as follows. An initial random partition into k clusters is performed on the data. The centroids (mean vectors) of each cluster are computed, and each object is re-assigned to the cluster with the nearest centroid. The centroids are

[5] Some clustering algorithms can also handle nominal features. We only use continuous features.

recomputed, and the process is repeated until no further changes take place, or a pre-specified number of times. A weakness of k-means is the fact that the initial assignment of items to clusters greatly affects the resulting structure. A common solution is to repeat the experiment several times, with different initial random assignments, and to adopt the solution that better satisfies the clustering criterion. In the k-means algorithm, the criterion used is to minimise the overall distance from objects to their centroids, which favours globular cluster structures.

Parameter	**Value**
clustering method	partitional, flat
similarity measure	cosinus
clustering criterion	minimise $\sum_{i=1}^{n} \cos(i, C_i)$
iterations	20
clusterisations	25
number of clusters (k)	3 and 5
number of features	6 and 36
feature values	standardised

Table 5.5: *Parameters for experiment A.*

The parameters of the experiment are summarised in Table 5.5. The three first parameters correspond to the definition of the k-means algorithm. [6] The *iterations* parameter specifies how many times the process of centroid computation and assignment to cluster is repeated within each clustering trial. The *clusterisations* parameter specifies how many times the whole clustering process is repeated. Parameters *iterations* and *clusterisations* were set to 20 and 25 respectively because they were shown to yield stable solutions.

We will discuss the solutions in 3 and 5 clusters (parameter *number of clusters*) because they best correspond to our intended classification. We have three main classes (intensional, qualitative, and relational) and a total of five classes (main classes plus polysemous classes: intensional-qualitative and qualitative-relational). The number of features corresponds to the number of features for the semantic and the distributional models of the data discussed in Section 5.1.2.

Finally, as can be seen in Table 5.5, feature values were not represented as simple proportions, but as standardised values, so that all features have mean 0 and standard deviation 1. This representation is achieved using the z-score formula shown in equation (5.1), where $\bar{x}$ is the mean of all feature values $x_1, \ldots, x_n$, and sd_x their standard deviation.

$$z_i = \frac{x_i - \bar{x}}{sd_x} \tag{5.1}$$

In clustering, features with higher mean and standard deviation values tend to dominate over more sparse features. Standardisation smooths the differences in the strengths of features. We experimented with raw and standardised values, and the most interpretable results were obtained with standardised values (although, again, the differences were not large). Therefore, we will only discuss results obtained with this parametrisation.

[6] In the formula for the clustering criterion, n is the total number of objects, and C_i is the centroid of the cluster for object i.

5.1.4 Results

5.1.4.1 Classes in clusters

The contingency tables of the clustering results with 3 clusters are depicted in Table 5.6. Rows are classes, named by their first letters (see the full legend in footnote 3, page 106). Columns are clusters, labeled with the cluster number provided by CLUTO. The ordering of the cluster numbers corresponds to the quality of the cluster, measured in terms of the clustering criterion chosen. 0 represents the cluster with the highest quality.

In each cell C_{ij} of Table 5.6, the number of adjectives of class i that are put in cluster j by the algorithm is shown. [7] Row $Total_{GS}$ contains the number Gold Standard lemmata that end up in each cluster. Row $Total_{cl}$ represents the total number of lemmata in each cluster, irrespectively of whether they belong to the Gold Standard or not. Recall from Section 5.1.1 that we cluster the whole set of 3,521 adjectives with more than 10 occurences in the corpus, although we only analyse the classification of the 101 lemmata that have been previously labeled.

	A: Sem.			**B: Distr.**			
Cl.	*0*	*1*	*2*	*0*	*1*	*2*	*Total*
I	0	0	**2**	0	**2**	0	*2*
IQ	0	0	**1**	0	**1**	0	*1*
Q	4	13	**35**	10	**37**	5	*52*
QR	3	5	3	7	2	2	11
R	**21**	13	1	**20**	5	10	*35*
$Total_{GS}$	*28*	*31*	*42*	*37*	*47*	*17*	*101*
$Total_{cl}$	*834*	*1287*	*1400*	*1234*	*1754*	*533*	*3521*

Table 5.6: *Experiment A: 3-way solution contingency tables.*

A striking feature of Table 5.6 is that results in each subtable (A and B) are very similar. In both solutions, the following can be observed:

- there is a cluster (labeled 0) that contains the majority of relational adjectives in the Gold Standard. This is the most compact cluster according to the clustering criterion.
- another cluster (2 in solution A, 1 in solution B) contains the majority of qualitative adjectives in the Gold Standard, as well as all intensional and IQ adjectives.
- the remaining cluster contains a mixture of qualitative and relational adjectives in both solutions.
- adjectives that are polysemous between a qualitative and a relational reading (QR) are scattered through all the clusters, although they show a tendency to be ascribed to the relational cluster in solution B.

The contingency table comparing the solutions obtained with semantic and distributional features, shown in Table 5.7, confirms that the results bear a close resemblance. Based on the number of objects shared, an equivalence between clusters can be established (0-0, 1-2, 2-1). The corresponding cells are boldfaced in the table.

[7] Note that, as the contingency tables of the semantic and distributional features have been collapsed into a single

		Distr			
		0	*1*	*2*	*total*
	0	**602**	47	185	*834*
Sem	*1*	595	396	**296**	*1287*
	2	37	**1311**	52	*1400*
	total	*1234*	*1754*	*533*	*3521*

Table 5.7: *Contingency table comparing semantic and distributional solutions.*

The main difference between the two types of features seems to be that in the solution obtained with theoretically features the clusters are "purer". Only 7 non-relational adjectives end up in the relational cluster, while for distributional features there are 17 non-relational adjectives in the relational cluster.

The 5-way solutions, depicted in Table 5.8, show more differences across feature type. We could expect the hybrid, Q/R cluster, to split into two pure Q and R clusters in the 5-way solution. What we find instead is that the hybrid clusters persist (cluster 0 in solution C, 0 and 1 in solution D), and that two further small clusters are created in each solution.

In the semantic solution (subtable C), there is in addition a relational cluster (cluster 1) and a qualitative cluster (cluster 2). The other two clusters seem to be subpartitions within the qualitative class. In the distributional solution (subtable D), the qualitative cluster is still to be found (cluster 2), but the relational cluster is lost: relational adjectives are scattered through all clusters, and tend to concentrate in clusters 0 and 1.

	C: Sem.					**D: Distr.**					
Cl.	*0*	*1*	*2*	*3*	*4*	*0*	*1*	*2*	*3*	*4*	*Total*
I	0	0	**2**	0	0	0	0	**2**	0	0	*2*
IQ	0	0	**1**	0	0	0	0	**1**	0	0	*1*
Q	7	4	**35**	4	2	3	7	**37**	2	3	*52*
QR	5	3	3	0	0	6	1	2	1	1	11
R	12	**21**	1	0	1	11	9	5	7	3	*35*
$Total_{GS}$	24	28	42	4	3	20	17	47	10	7	*101*
$Total_{cl}$	857	854	1462	156	192	828	406	1754	275	258	*3521*

Table 5.8: *Experiment A: 5-way solution contingency tables.*

In the next section, we will see that clusters 3 and 4 are the least clearly interpretable in both solutions. They are also the poorer clusters according to the clustering criterion. A possible explanation is that they contain the least frequent adjectives, as shown in Table 5.9, so that the clustering algorithm just does not have enough information for these items. One-way ANOVA of frequency against cluster label confirms that the difference in frequency is significant (data also Table 5.10).

The data in Tables 5.6 and 5.8 are shown as barplots in Figure 5.2 for clarity. The upper graphics (A and B) depict the 3-way solutions using semantic and distributional features, respectively.

table, the column named *total* represents the row sum of each subtable (the number of items per class is constant).

	0	**1**	**2**	**3**	**4**
Semantic	141 ±307	118 ±303	133 ±446	61±126	58±142
Distributional	98 ±249	119 ±281	148 ±482	108±256	60±135

Table 5.9: *Frequency of adjectives across clusters.*

	df	**F**	**p**
semantic	4;905.3	30.7	$< 10^{-15}$
distributional	4;674	18.7	$< 10^{-13}$

Table 5.10: *Results for ANOVA of frequency against cluster label.*

The lower graphics (C and D) depict the 5-way solutions. The labels A, B, C, and D match the data with the same labels in Tables 5.6 and 5.8. Each bar is a cluster (cluster label below the bar), and each colour represents a class.

In Figure 5.2, the similarity between the solutions is graphically evident, as is the purer correspondance between clusters and classes when the solutions are obtained with semantic features as opposed to distributional features, particularly in the 5-way solution. Also note that in the distributional solution, the tendency is for a large cluster (number 2 in graphic D) to concentrate most objects and the remaining clusters to be comparatively small. In contrast, the semantic solution has three large clusters and two extremely small ones, suggesting that the discriminative power of semantic features is higher than that of distributional features. We will provide an explanation for this difference in Section 5.1.4.2.

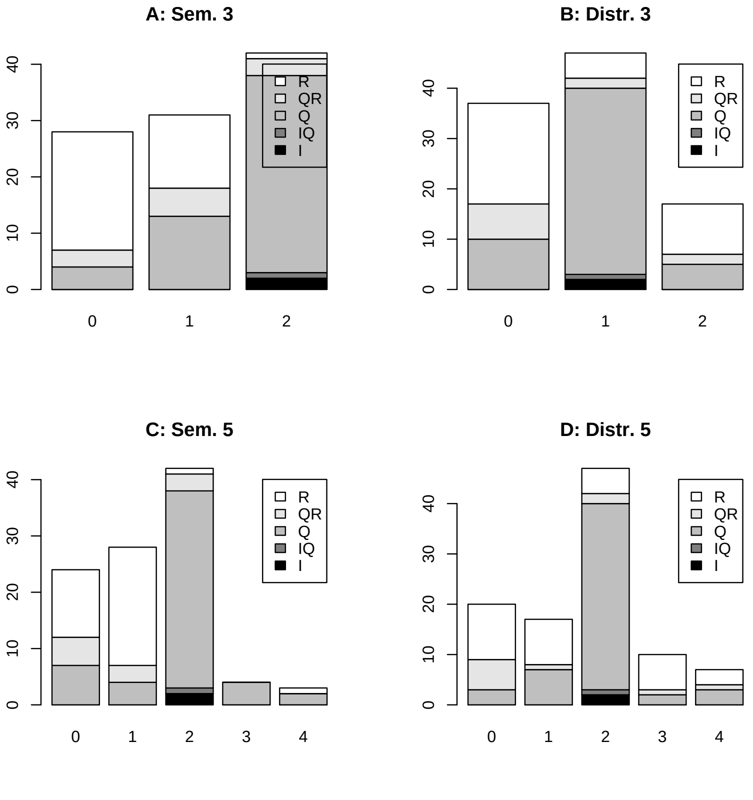


Figure 5.2: *Experiment A: results.*

We have performed a qualitative evaluation based on a comparison between the Gold Standard and the clustering solution. However, we have not provided a numerical evaluation. Evaluation of clustering is very problematic when there is no one-to-one correspondence between classes and clusters (Hatzivassiloglou and McKeown, 1993), as is our case. Schulte im Walde (2006) provides a thorough discussion of this issue and proposes different metrics and types of evaluation. Because of the exploratory nature of the experiment, and because the classification will be changed according to its results, we will defer numerical evaluation until the next experiment, explained in Section 5.2.

5.1.4.2 Feature analysis

We now turn to feature analysis. We concentrate in the 5-way solutions because they match our number of targeted classes. Figure 5.3 shows the feature value distribution in the 5-way solution using semantic features. If the solution matched the human classification, its shape should be very similar to Figure 5.1 (page 105), although Figure 5.3 covers the distribution of all 3,521 adjectives, and Figure 5.1 only that of the 101 labeled adjectives of the Gold Standard. Some resemblance is observed, but no one-to-one correspondence.

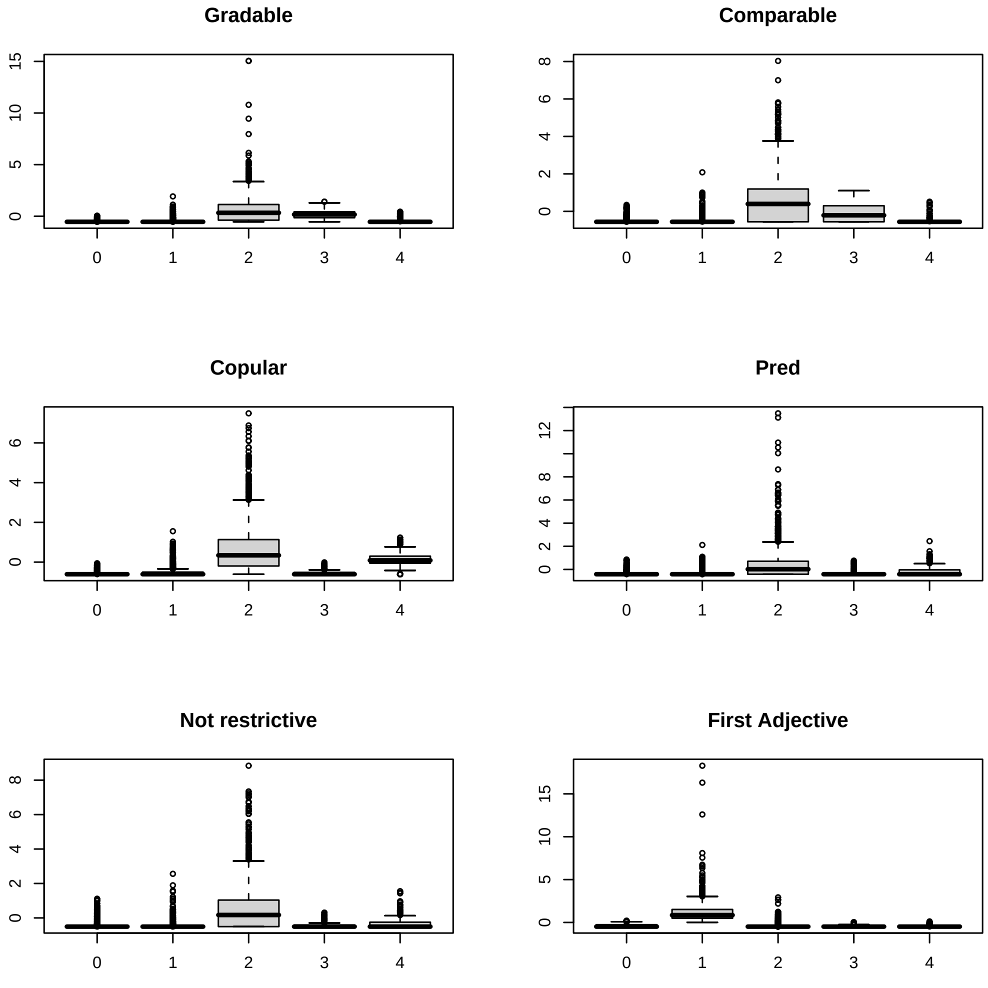


Figure 5.3: *Feature value distribution across clusters (semantic features).*

The feature distribution is in accordance with the analysis developed in the previous section. Cluster 0, the hybrid cluster, is negatively defined: it has mean values below the grand 0 mean for all features. Clearly, the features chosen fail to characterise the lemmata it contains.

Cluster 1, the relational cluster, has negative (thus lower than the mean) mean values for all features except for feature *adjacent*, for which it has a mean value of 1.3, that is, more than one standard deviation above the mean. The prediction outlined in page 106 matches this characteristic, indicating that relational adjectives tend to occur closer to the head noun when more than one adjective modifies it.

For cluster 2, the qualitative cluster, the situation is the reverse, as predicted: it presents positive (above mean) values for all features except for *adjacent*. Thus, the typical qualitative adjective is gradable, comparable, predicative, can be used for non-restrictive modification, and occurs after other modifying adjectives. Note, however, that the feature value distributions are all Zipfean and in some cases (such as clusters 1 and 2) there is wide within-cluster variability, which is clear from the quantity of points (outliers) above the upper tail of the boxplots. This means that for most qualitative adjectives, their characteristic features (gradability, predicativity, etc.) have low values, while for a few there is abundant evidence. This distribution favours confusion of relational and qualitative adjectives.

As for clusters 3 and 4, which are much smaller than clusters 0-2 and according to our Gold Standard are subdivisions within the qualitative class, they contain qualitative adjectives that exhibit a low predicativity (cluster 3; see values for feature *copular* and *pred*) or a low gradability (cluster 4; note values for features *gradable* and *comparable*). However, recall from Section 5.1.4.1 (Table 5.9) that adjectives in these two clusters have a lower frequency than in the remaining clusters, suggesting that data sparseness negatively affects clustering results.

For the distributional solution, 36 features were used, which is too high a number for a graphical representation in the fashion of Figure 5.3. Instead, Table 5.11 shows, for each cluster, the 3 features with the highest and lowest relative values, respectively. [8] Their mean and standard deviation values are recorded under each feature label (standard deviation values are indicated with the ± sign and in a smaller font).

Note that the standardised feature representation makes this type of exploration easier than percentage representation: features with large positive values indicate that the objects in the relevant cluster occur in the distributional context with higher frequency than the remaining objects, and the reverse is true for features with large negative values. If the distribution of adjectives into clusters were random, all feature values would be around the grand mean, that is, around zero.

The last column of Table 5.11 records the equivalence for each cluster in the semantic solution, based on the number of objects shared. [9]

Cluster 0, containing mostly relational and some QR adjectives, has high positive values for the default position of adjectives in Catalan (*-1 noun, -2 determiner*), and low values for features

[8] Abbreviations:

adj.:	adjective
det.:	determiner
conj.:	conjunction
prep.:	preposition
punct.:	punctuation

[9] The contingency table of the two solutions is shown below. Rows correspond to semantic features, columns to distributional features. The equivalences recorded in Table 5.11 are boldfaced.

Cl.	**Highest values**			**Lowest values**			**S.**
0	*-2 det.,*	*-1 noun,*	*+1 verb*	*-2 verb,*	*-1 adverb,*	*-2 noun*	1
	1.1 ±0.61	1 ±0.51	0.86 ±1.2	-0.66 ±0.42	-0.64 ±0.32	-0.62 ±0.53	
1	*+1 prep.,*	*+2 det.,*	*-2 det.*	*-1 adverb,*	*-2 verb,*	*-1 punct.*	0
	0.93 ±0.94	0.88 ±1	0.81 ±0.67	-0.56 ±0.42	-0.53 ±0.54	-0.43 ±0.48	
2	*-2 verb,*	*-1 adverb,*	*-1 verb*	*-1 noun,*	*-2 det.,*	*+1 verb*	2
	0.56 ±1.1	0.55 ±1.1	0.49 ±1.2	-0.8 ±0.62	-0.73 ±0.61	-0.39 ±0.68	
3	*-2 adj.,*	*-1 conj.,*	*+2 adj.*	*-2 noun,*	*-2 verb,*	*-1 verb*	1
	1.1 ±1.1	0.64 ±1.1	0.82 ±1.4	-0.49 ±0.6	-0.46 ±0.6	-0.45 ±0.6	
4	*-2 prep.,*	*+1 punct.,*	*-1 noun*	*+1 prep.,*	*-1 adverb,*	*-2 noun*	0
	1.2 ±1.4	0.56 ±0.93	0.51 ±0.65	-0.51 ±0.58	-0.41 ±0.54	-0.4 ±0.79	

Table 5.11: *Highest and lowest valued distributional features across clusters.*

typical of qualitative adjectives: predicative contexts (*-2 verb*), gradability (*-1 adverb*), and noun modifier with some element (e.g., other adjectives) in between (*-2 noun*). This means that adjectives in this cluster occur almost only rigidly attached to the noun, as is expected for relational adjectives.

Cluster 1, equivalent to the "hybrid" cluster in the semantic solution, also exhibits low values for features typical of qualitative adjectives (*-1 adverb*, *-2 verb*), the main difference being that lemmata in this cluster tend to appear in complex NPs with prepositional adjectives (high values for *+1 preposition, +2 determiner*). Furthermore, 9 of the 17 Gold Standard adjectives in this cluster are deverbal. This suggests that some deverbal adjectives behave neither like qualitative adjectives nor like relational adjectives. The clustering elicits deverbal adjectives, which might tend to have more complex argument structures, and in particular to bear complements. This issue will arise again in Section 5.2.

Cluster 2, the qualitative cluster, is almost symmetric to cluster 0: it has values below the mean for the "default" features (*-1 noun* and *-2 determiner*), and higher values for typically qualitative features (*-2 verb*, *-1 adverb*, *-1 verb*). This value points to the greater syntactic flexibility available for qualitative adjectives, as discussed in Chapter 3: if they appear less tightly attached to the noun, they can appear in other contexts, mostly in gradable and predicative contexts.

Recall from Section 5.1.4.1 that cluster 2 is the one that concentrates almost all data. This cluster is dominated by features *-1 noun* and *-2 determiner* (the values with largest absolute mean values). In this experiment, we did not perfom feature selection for the distributional solution, with the reasoning that distributional features represent a neutral (if simple) representation of an adjectives' linguistic behaviour. The problem is that, despite the standardisation performed on feature values, some features are still very dominant, most notably the first ones in Table

		Distr					
		0	1	2	3	4	*Total*
Sem	0	276	**203**	188	85	**105**	*857*
	1	**497**	120	47	**119**	71	*854*
	2	15	31	**1353**	27	36	*1462*
	3	17	21	78	22	18	*156*
	4	23	31	88	22	28	*192*
	Total	*828*	*406*	*1754*	*275*	*258*	*3521*

5.3, which include *-1 noun* and *-2 determiner*. In addition, these two features have a correlation coefficient of 0.80, highly significant (t(3519)=80.5, p $< 10^{-15}$). These two factors cause distributional features to have less discriminant power than semantic features, which is even clearer in solutions with a higher number of clusters. To avoid this effect, in the second set of unsupervised experiments, explained in Section 5.2, feature selection was performed to test whether it improved interpretability of the clusters.

As for the smaller clusters, 3 and 4, they do not have a clear interpretation in terms of the parameters discussed in Chapter 3. Adjectives in cluster 3 seem to be relational adjectives (low values for *-2 noun, -2 verb, -1 verb*) that appear in coordinating constructions (high values for *-2 adjective, -1 conjunction, +2 adjective*).

Cluster 4 is even less clear. Exploration of the adjectival lemmata in this cluster did not yield a compact characterisation, but a few coherent subgroups. First, it contains quite a large number of adjectives related to medicine (12 lemmata), among others *antisèptic*, *eritematós*, *mamari*, and to other scientific activities (22 lemmata), such as *amònic*, *binari*, *gasós*, and *palatal*. It makes sense for these adjectives to occur in bare NPs, that is, NPs with no determiner (high values for *-2 preposition* and *-1 noun*), because they are typically involved in NPs expressing substance or another kind of classification. However, this cluster also contains some colour adjectives (8 lemmata), such as *blanc*, *blau*, and *castany* (‘white’, ‘blue’, ‘chestnut brown’), and some quite ordinary qualitative adjectives, such as *brusc*, *corb*, o *invers* (‘sudden’, ‘crooked’, ‘inverse’). Most of the qualitative adjectives in the cluster, however, seem to be nongradable (which accounts for low value in *-1 adverb*).

The caveats raised for semantic clusters 3 and 4 apply here. Clusters 3 and 4 are the less compact (according to the clustering criterion) clusters, so it is should be expected that they are the hardest to interpret. In addition, adjectives in these two clusters have a lower frequency (see Table 5.9 in Section 5.1.4.1) so are presumably affected by data sparseness. Also note that the standard deviation values depicted in the table are very large, often larger than the mean value itself (two thirds of the value). This indicates a high variability of the data.

Beyond the interpretation of each particular cluster, there is a striking fact about Table 5.11: among the 36 distributional features used, only a few recurrently appear in the characterisation of the clusters in terms of their highest and lowest values. Moreover, it turns out that most of the recurrent features are highly correlated with one or more of the semantic features. That explains the similarity between the solutions obtained with the two kinds of features. It also provides empirical support for the decisions involved in the definition of semantic features: if the syntactic context is blindly provided to the clustering algorithm, it selects as most relevant for establishing groups of adjectives most of the properties discussed in Chapter 3.

Those features that appear more than once in Table 5.11 are listed in Table 5.12. The second column records the number of times they appear in Table 5.11, and the remaining columns specify the Pearson correlation coefficient with each of the semantic features. The correlation coefficient is computed taking into account all adjectives clustered, that is, 3,521. [10]

[10] Abbreviations with respect to the nomenclature used in Table 5.1:

grad.:	gradable
comp.:	comparable
not restr.:	not restrictive
adjac.:	adjacent

Note that most of the features in Table 5.12 correspond to the left context (positions -1 and -2). This will be taken into account in the design of the features for experiment B (Section 5.2).

Feature	#	Grad.	Comp.	Copular	Pred	Not restr.	Adjac.
-1 adverb	4	**0.67**	**0.75**	**0.47**	0.19	0.16	−0.26
-2 verb	4	**0.42**	0.29	**0.43**	**0.40**	0.27	−0.26
-2 determiner	3	−**0.40**	−0.35	−**0.47**	−0.33	−0.32	0.33
-1 noun	3	−**0.43**	−**0.42**	−**0.50**	−0.36	−0.39	**0.41**
-2 noun	3	0.27	0.33	0.24	0.10	0.13	−0.24
-1 verb	2	0.18	0.10	**0.71**	**0.49**	0.05	−0.25
+1 verb	2	−0.20	−0.19	−0.19	−0.16	−0.21	0.13
+1 preposition	2	−0.03	−0.02	0.12	0.02	−0.18	−0.10

Table 5.12: *Correlation between distributional and semantic features.*

All correlation values are highly significant (p < 0.01), due to the large number of objects available for comparison. However, some are particularly large. Absolute values ≥ 0.40 have been highlighted in the table.

Feature *-1 adverb*, as would be expected, is highly correlated with features *gradable* and *comparable*. However, it is also highly correlated with feature *copular*. Similarly, feature *-2 verb* is highly correlated with features *copular* and *pred* (to be expected, as it is a shallow cue of a predicative use of the adjective), but also with *gradable*. These two pieces of evidence support the relationship between predicativity and gradability expressed in Chapter 3: adjectives that frequently occur in predicative contexts tend to be gradable, although there are many exceptions. The fact that these features are prominent in qualitative clusters in all solutions also supports the class and its characterisation.

Note that *-1 adverb* and *-2 verb* are positively correlated with all semantic features except for *adjacent*, the only feature that was designed specifically to charaterise relational adjectives. The same can be said of *-2 noun* and *-1 verb* (the latter is highly correlated with *copular* and *pred*, as could be expected). The situation is the reverse for features *-2 determiner* and *-1 noun*, which are negatively correlated with all features except for *adjacent*. The most natural shallow correlate for *adjacent*, namely, *+1 adjective*, is not present in Table 5.12, presumably because it is a too sparse feature. However, features *-2 determiner* and *-1 noun* convey similar information.

The last two features, which only appear twice in Table 5.11, are not highly correlated with any of the theoretical features. Feature *+1 verb* could point to subject-modifying adjectives; the role of the syntactic function of the modified heads will be explored in Chapter 6. As for feature *+1 preposition*, it should be typical of complement-bearing adjectives, and will be further discussed in the experiments reported in Section 5.2.

Five of the eight features depicted in Table 5.12 are among the 10 most frequent distributional features, listed in Table 5.3 above. It could be argued that this explains their relevance in the clustering results: if they are the most frequent features, the richest evidence will be available for them. However, recall that the semantic features were defined independently of the distributional features. The coincidence between the two levels of representation is what Table 5.12 highlights. The fact that distributional features in Table 5.12 are among the most frequent features indicates that they are the most representative contexts for adjectives – which adds support to the present definition of the semantic features.

5.1.5 Discussion

The analysis of the results allows us to draw some conclusions with respect to the semantic classification of adjectives, which was one of the main goals of the experiments reported in this section.

Both the semantic solution and the distributional solution provide empirical support for the qualitative and relational classes, as is particularly evident in the 3-way solution. Intensional and IQ adjectives are systematically grouped together with qualitative adjectives in all solutions, so that they do not have syntactic characteristics that are strong enough to differentiate them. In this respect, note that the main syntactic characteristic of most intensional adjectives, pre-nominal position, is shared by a large number of qualitative adjectives (although for most intensional adjectives this position is obligatory and for qualitative adjectives in general it is not).

As for QR (polysemous between a qualitative and a relational reading) adjectives, they are spread through all the clusters in all solutions: they are not identified as a homogeneous group (they are not grouped together) nor as distinct from the rest (they are in clusters that contain other kinds of adjectives). These adjectives usually have feature values in between those of the main classes, but the differences are not strong enough to motivate a separate cluster.

The present approach is clearly not adequate to model polysemy. The problem could lie in the experimental procedure, either in the modeling of the data or the algorithm chosen. However, a more plausible cause is that polysemous adjectives do not have a homogeneous, differentiated profile. Most adjectives are used predominantly in one of their senses, corresponding to one of the classes. For instance, *irònic* ('ironic'), classified as QR, is mainly used as qualitative in the corpus. Accordingly, it always appears in the qualitative clusters. Conversely, *militar* ('military'), also classified as QR, is mostly used as relational, and is consistently assigned to one of the relational clusters. Therefore, the treatment of polysemy in terms of additional, "polysemous" classes to be separately acquired, is not adequate. We will test an alternative conceptualisation and experimental design in Chapter 6.

What about the "hybrid" cluster? This cluster seems to be coherent and stable: it appears in all the solutions examined, and is quite compact in terms of the clustering criterion (it is higher ranked by CLUTO in terms of cluster label). This is a good candidate to signal problems in the proposed classification. A comparison of the classifications of the human judges and the clustering solutions shows that most of the adjectives that are problematic for humans (i.e., are classified differently by different judges) are in the hybrid cluster of the 5-way semantic solution. Conversely, most adjectives in this cluster are problematic.

The Gold Standard lemmata assigned to cluster 0 are the following (problematic adjectives underlined): *accidental* ('accidental'), *alemany* ('German'), *alfabètic* ('alphabetical'), *anticlerical* ('anticlerical'), *caracurt* ('short-faced'), *celest* ('celestial'), *diversificador* ('diversifying'), *femení* ('feminine'), *gradual* ('gradual'), *indicador* ('indicating'), *menorquí* ('Menorcan'), *negatiu* ('negative'), *parlant* ('speaking'), *preescolar* ('pre-school'), *protector* ('protecting/protective'), *salvador* ('saviour'), *sobrenatural* ('supernatural'), *sud-africà* ('Sudafrican'), *triomfal* ('triumphal'), *tàctil* ('tactile'), *valencianoparlant* ('Valencian-speaking'), *ventral* ('ventral'), *veterinari* ('veterinarian'), *xinès* ('Chinese').

17 out of the 24 Gold Standard adjectives in this cluster (70.1%) are problematic for humans. In the qualitative cluster (cluster 2), only 10 out of 42 lemmata (23.8%) are problematic, that is, are not assigned to the same class by all judges. Two kinds of adjectives strike among

problematic adjectives: nationality-denoting adjectives (*alemany, menorquí, sud-africà, xinès*), and deverbal adjectives (*indicador, parlant, protector, salvador*). These two kinds of adjectives do not fit into the classification as it was proposed to the human judges.

Nationality-denoting adjectives, as discussed in Section 3.6.4, can act as predicates of copular sentences in a much more natural way than typical relational adjectives, and seem to be ambiguous between a relational and a qualitative reading in their semantics. This kind of adjectives is treated as polysemous in the Gold Standard for final experiments reported in Chapter 6, as has been explained in Section 4.5.

The problem for deverbal adjectives is similar. They are clearly neither relational (no object or nominal to relate to) nor intensional (they share neither entailment patterns nor syntactic behaviour). However, they are also not typically qualitative. For instance, while their use as predicates is perfectly natural (example (5.2a)), they typically do not exhibit the entailment behaviour of qualitative adjectives (examples (5.2a-5.2a)). Also, many deverbal adjectives trigger a clear relationship to an event (a protecting event, in the case of *protector*). Therefore, we decided to introduce the event-related class as a further class in the classification.

(5.2) a. El Joan és protector
The Joan is protective

‘Joan is protective’

b. Serra . . . Era soci protector de l’Associació de concerts
Serra . . . was associate protecting of the-Association of concerts

‘Serra was a protecting associate of the Association of concerts’

c. ⊭ ?#Serra era protector

The cluster analysis as performed in this Section has allowed us to identify two kinds of adjectives that do not fit into the classification as it was initially designed. It has thus provided empirical evidence for the need of refining the criteria (classifying nationality-denoting and similar adjectives as polysemous between a relational and a qualitative reading) or even the classification itself (proposing the event-related class). The latter decision is backed up by some theoretical proposals, such as Ontological Semantics, as discussed in Chapter 3.

As a result of the experiments presented in this Section, we modify our proposal of semantic classification for adjectives. The event-related class is introduced for the reasons just discussed, and the intensional class is removed.

As discussed in Chapter 3, the intensional class is very small[11] and its members are not homogeneous, neither in their semantics nor in their syntax (see the discussion about modal adjectives in Section 3.3.2.2). For NLP purposes, it can be manually handled, given its size.

In addition, recall from Section 3.6 that we want our classification to be consistent, in the sense that its parameters should be homogeneous for all classes. For qualitative, relational, and event-related adjectives, this parameter is the ontological type of denotation: qualitative adjectives denote attributes (in Raskin and Nirenburg’s (1998) terminology), relational adjectives denote relationships to objects, and event-related adjectives denote relationships to events. The defining trait of the intensional class is on the contrary its behaviour with respect to entailment

[11] Recall that two prototypical intensional adjectives had to be manually introduced in a post-hoc fashion, as they were not represented in the randomly chosen Gold Standard.

patterns, or the fact that they denote second order properties. Removing this class allows us to improve consistency in the classification.

The class labels are changed to make this shift in classification parameter clear: in the remaining of this thesis, the classification pursued will distinguish between **basic** (formerly qualitative), **object-related** (formerly relational), and **event-related** adjectives. The change in denomination includes ontological terms for the class names, closely following the terminology in Raskin and Nirenburg (1998).

An exception is the basic class, named *scalar* in Raskin and Nirenburg (1998). As discussed in Chapter 3, it is not clear that all adjectives included in their scalar class can be organised in terms of a lexically defined scale. Terms like *property-based* or *attribute-based,* also used in Raskin and Nirenburg (1998), induce confusion with the terms *property* in formal semantics (all adjectives denote properties) and *attribute* in several uses (e.g., as synonymous to “syntactic modifier”). With the rationale that the most prototypical adjectives are included in this class, we choose the term *basic*. However, of course, the label does not *per se* define or change the content of the class.

We finish this discussion with a summary of two further aspects that have been raised in the course of the experiment. One goal of this experiment, as stated at the beginning of this chapter, was to compare the modelling of adjectives in terms of theoretically biased features (termed *semantic* features in the explanation) with a neutral model in terms of the whole distributional context, represented as POS unigrams. It came as a surprise that the results obtained with the two representations bear a close resemblance. In the analysis, we have argued that the most influential features in the resulting clusters mainly correspond to those defined according to theoretical considerations. This result supports the present definition of theoretical features.

Finally, the analysis of the results has also allowed us to identify some aspects of the experiment that should be improved, among them the following two. First, the minimum of 10 occurences is clearly too low a threshold. In subsequent experiments explained in Section 5.2 and Chapter 6, the threshold is raised to 50 occurences. Second, some features are too dominating, causing distributional features to have less discriminating power than semantic features, so that items tend to concentrate in a single cluster. The correlation between features representing different positions of the context seems to strengthen this effect. Attribute selection and combination (considering more than one position of the context in a single feature) will be performed in the remaining experiments so as to avoid this effect.

We now turn to discussing a clustering experiment performed to analyse the new classification proposal.

5.2 Experiment B: testing the new classification

The experiment reported in Section 5.1 provided some evidence as to what features could be relevant. It also showed that simple distributional features defined in terms of n-gram distribution yield very similar results to those obtained with carefully designed features. Finally, it showed that using distributional features indiscriminately obscured results, because the most dominant features caused most objects to be clustered together.

In the experiment reported in this section, we further pursue the use of distributional features. Because we use clustering as an Exploratory Data Analysis tool, so as to gain insight into the characteristics of the new classification proposal, it makes sense to use a neutral representation of the linguistic behaviour of adjectives, as opposed to theoretically biased features.

However, given the problems encountered in the previous experiment, we do not use the whole bunch of distributional features, but perform feature selection. Thus, we perform the experiment in two steps: first, we analyse the feature distribution of the different classes, obtained from a set of manually labeled adjectives. We choose the features that best characterise each class, and perform clustering experiments using only the selected features. The results are analysed with a different set of manually labeled adjectives, so as to avoid overfitting.

5.2.1 Classification and Gold Standard

The classification distinguishes between basic, event-related (*event* for short), and object-related (*object* for short) adjectives.

As reported in Chapter 4, we built a small Gold Standard for the purposes of this experiment, corresponding to Gold Standard B in Table 4.25 (page 102). In addition, as stated in the introduction to this section, we used another subset of manually annotated data for feature selection purposes (*tuning subset* from now on).

The Gold Standard consists of 80 lemmata and was classified by 3 judges along two parameters. The 3 classifications were subsequently merged into a single one by the author of this thesis to analyse the clustering results.

The first parameter is the semantic classification, distinguishing between basic, event and object adjectives. Recall that the judges were allowed to assign a lemma to a second class in case of polysemy, and that agreement scores for polysemy judgments were not significant at all. Therefore, for this experiment we only consider the main (first) class assigned by a judge, and the acquisition of polysemy is deferred until the supervised experiments explained in Chapter 6. The mean K score for the main class of an adjective is 0.74, which can be safely accepted for academic purposes.

The second parameter for classification within the Gold Standard is the distinction between *unary* and *binary* adjectives, depending on the number of arguments of a particular adjective. Adjectives usually have a single argument (the head noun), and are thus *unary*. However, some adjectives have two arguments (are *binary*), and the second argument is syntactically realised as a PP or clause complement. An example is *gelós* ('jealous'), as in *El Joan està gelós de la Maria* ('Joan is jealous of Maria'). Recall from Section 3.1 that adjectives with more than one complement are extremely rare (see examples in (3.3b), page 21), so that we only considered the unary vs. binary distinctions. The judges had a mean K value of 0.72 for this task. This parameter will be relevant for the analysis of results.

The tuning subset consists of 100 adjectives, randomly chosen and manually classified by the author of the thesis.

5.2.2 Features

The discussion in the previous chapter has shown that many features correlate, because the left and right contexts of adjectives are not independent. Accounting for both contexts at the same time in the feature definition should improve descriptive coverage, although it also boosts the number of features, which raises questions of data sparseness. The left context has been shown in Section 5.1.4.2 to be more significant than the right context: the features with highest and lowest values in the resulting clusters (Table 5.11, page 5.11) were mainly those corresponding to the part of speech of words preceding the target adjective.

A powerful model to account for these facts would be a 5-gram model, containing three words to the left and one to the right of the adjective. However, a 14.5 million corpus such as the one we use here does not provide enough evidence for a 5-gram representation. We therefore use bigram pairs instead: in a 5-word window, the first two tags form a feature and the second two tags another feature.

We also include different distinctions than those provided in the original part of speech tags, so as to make them fit our task better. The POS tags we use for representation are obtained by combining the information in the first and second level of the original tagset. They are listed in Table 5.13. [12]

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 (marked with an asterisc in Table 5.13) were assigned the tag *empty*.

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 5.13: *Tags used in the bigram representation.*

Example (5.3) shows the representation that would be obtained for the sentence we have reviewed in the first experiment (example (5.1), page 5.1). The target adjective, *fix*, is in bold face, and the relevant word window is in italics. Negative numbers indicate positions to the left, positive ones positions to the right.

(5.3) a. Els instints domèstics, per dir-ho així, són del *cert molt menys* **fixos**, o
The instincts domestic, to say-it so, are of-the true very less fixed, or
invariables, que els instints naturals
invariable, than the instincts natural

‘Domestic instincts, to put it this way, are really much less fixed, or invariable, than natural instincts.’

b. -3aj-2av, -1av+1co

The representation for sentence (5.3) states that the first element of the n-gram (-3; third word to the left of the adjective) is an adjective, the second element is an adverb, the third one (-1; word preceding the adjective) is also an adverb, and the fifth one (+1; word following the adjective) is a coordinating element. The two first elements form a feature (-3xx-2yy) and the two remaining elements form another feature (-1ww+1zz).

[12] Clause delimiters are punctuation marks other than commata, relative pronouns and subordinating conjunctions. Coordinating elements are commata 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.

This representation schema produced a total of 240 different feature (bigram) types, 164 of which had a prior probability < 0.001 and were discarded. The number of remaining features, 76, made it impossible to perform automatic feature selection, for it almost matched the number of objects in the tuning subset (100). We attempted at selecting features according to the p value obtained with a statistical test, individually applied to each feature. However, we faced difficulties with establishing a reasonable threshold, because p values varied a lot depending on the mean frequency of the feature.

For these reasons, we performed manual feature selection, on the basis of the exploration of the boxplots for the tuning subset. In the experiments explained in Chapter 5.2, we use a larger Gold Standard, which allows us to perform automatic feature selection.

5.2.3 Clustering parameters

The clustering parameters for this experiment are very similar to the ones used for Experiment A (see Section 5.1.3).

Again, we tested several clustering approaches and criterion functions available in CLUTO, and again, the overall picture of the results was quite robust across different parametrisations. In addition, two different combinations of features and feature value representations (raw or standardised proportions) were tested for each parameter. For clarity reasons, we will limit the discussion of the results to the 3 cluster solution obtained with 32 features and standardised feature representation.

Also note that, as in Experiment A, we clustered the whole set of adjectives that occurred at least 50 times in the corpus (totalling 2,283 lemmata), and analysed the results by comparison with the 80-unit Gold Standard.

5.2.4 Results

5.2.4.1 Evaluation

The contingency table comparing classes and clusters is depicted in Table 5.14.

		Clusters			
		0	1	2	*Total*
Classes	basic	9	**26**	4	*39*
	event	2	7	**7**	*16*
	object	**25**	0	0	*25*
	$Total_{GS}$	*36*	*33*	*11*	*80*
	$Total_{cl}$	949	638	696	*3521*

Table 5.14: *Experiment B: 3-way solution contingency tables.*

Table 5.14 shows that, in contrast with Experiment A, there is a clear correspondence between clusters and classes, in the sense that in each cluster there is a majority of lemmata of one of the classes. Thus, cluster 0 contains mostly object adjectives, cluster 1 basic adjectives, and cluster 2 event adjectives.

The correspondence between clusters and classes makes numerical evaluation straightforward,

	p_o	K
baseline	0.49	10^{-15}
clustering	0.73	0.56
human	0.89	0.83

Table 5.15: *Experiment B: Evaluation.*

because cluster labels can be identified with one of the classes and numerical evaluation can proceed as usual with supervised techniques. The accuracy of the clustering algorithm (equivalent to the p_o agreement measure) is 0.73, that is, 73% of the lemmata were assigned by the algorithm to the expected cluster, given the equivalence just outlined. An adequate baseline accuracy for our task is that of assigning all lemmata to the most frequent class, namely, the basic class (39 out of the 80 lemmata in our Gold Standard are basic; see class distribution in the last column of Table 5.14). It results in 0.49 accuracy. The accuracy of the algorithm is almost 25% higher.

For comparison, it is useful to note that the K score of the clustering solution as compared with the Gold Standard is 0.56. Recall that mean inter-judge K is 0.74. The mean K value between each human judge and the Gold Standard obtained by merging the 3 classifications is 0.83 (range: 0.74-0.87), and the p_o is 0.89 (range: 0.83 to 0.93). The K value obtained with the clustering result is lower than human K values, but represents moderate agreement according to the scale proposed by Landis and Koch (1977). Table 5.15 summarises the results of the numerical evaluation for Experiment B. All values are obtained by comparing each of the classifications (baseline, clustering result, and human judges) with the Gold Standard. The last row corresponds to mean values averaged over the 3 judges. Note that the baseline K (obtained by comparing the Gold Standard with a uniform assignment to basic) is very near 0, as expected.

The class that receives less support in the cluster analysis is the event-related class, as half of the lemmata are grouped together with basic adjectives and half are in the event cluster. Our preliminary diagnostic, which will be confirmed by the experiments explained in Chapter 6, is that it is due to the lack of syntactic homogeneity of the event-related class.

A closer look at the lemmata in cluster 2 (the event cluster) reveals that it contains seven out of the eight binary adjectives in the Gold Standard, and only four unary ones. Recall from Section 5.2.1 that information about the number of arguments (distinguishing unary adjectives from binary ones, which have an additional argument realised as a complement) was encoded by the human judges. It seems, then, that what is being spotted in cluster 1 are binary, rather than event-related, adjectives. If we look at the morphological type, it turns out that six out of seven event adjectives in cluster 2 (against three out of seven in cluster 1) 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. Further discussion of this issue will be provided in Chapter 6.

Event adjectives do not form a homogeneous class with respect to the features used. In contrast, basic and object adjectives are quite clearly distinguished from each other and from event adjectives in the clustering solution depicted in Table 5.14, as was the case in Experiment A.

5.2.4.2 Feature analysis

As for the features that were most relevant for each cluster, listed in Table 5.16, they confirm the analysis just made. [13]

Cl.	Highest values			Lowest values		
0 (O)	*-3ey-**2dd***,	***-1cn**+1cd*,	***-1cn+1aj***	*-3ey-2ey*,	***-1av**+1cd*,	***-1av**+1co*
	0.98 ±0.77	0.91 ±0.71	0.85 ±1.2	-0.79 ±0.44	-0.62 ±0.33	-0.61 ±0.43
1 (E)	*-1co**+1pe***,	***-1ve+1pe***,	*-1cd**+1pe***	***-1cn**+1co*,	***-1cn**+1cd*	***-1cn**+1aj*
	1.22 ±1.5	1.19 ±1.6	1.11 ±1.6	-0.92 ±0.69	-0.79 ±0.70	-0.56 ±0.28
2 (B)	***-1av**+1cd*	***-1av+1co***,	***-1co**+1cn*	*-3ey-**2dd***,	***-1cn**+1ve*,	***-1cn**+1pe*
	0.66 ±1.1	0.65 ±1.1	0.54 ±1.3	-0.67 ±0.46	-0.53 ±0.47	-0.52 ±0.65

Table 5.16: *Highest and lowest valued bigram features across clusters.*

Note that most of the values in Table 5.16 correspond to the immediate context surrounding the adjective, thus supporting the claim that a small window is sufficient for the purposes of our task. In the supervised experiments explained in Chapter 6, only the left and right word of the adjective will be taken into account for the bigram representation.

Lemmata in cluster 0 (object adjectives) have high values for the expected "rigid" position, right after the noun (*-1cn*, preceded by common noun, with a determiner to its left, *-2dd*) and before any other adjective (*+1aj*, followed by adjective). They are further characterised by not being gradable (low value for features with *-1av*, preceded by adverb).

As for adjectives in cluster 1 (event adjectives), they are positively characterised by occuring before a preposition (*+1pe*), which is an indication of their bearing a complement. They also tend to occur in contexts other than the typical postnominal position (low values for *-1cn*), most notably in predicative position (high value for *-1ve*).

Finally, cluster 2 (basic adjectives) are characterised by being gradable (features with *-1av*), and for presenting a more flexible behaviour than the other two classes, as shown by their participation in coordinating constructions (*-1co*, *+1co*). Note that in this clustering solution, predicativity does not seem to play a major role in characterising basic adjectives. Although two features related to predicativity do present higher mean values for objects in cluster 2 than for objects in other clusters (*-1ve+1cd*: 0.31, *-1ve+1co*: 0.47), features regarding gradability and coordination have higher values.

Also note that the most compact class, both from the ranking assigned by CLUTO (cluster 0) and from the comparison with the Gold Standard, is that of object-related adjectives. It may seem surprising, as the most prototypical adjectives are in the basic class.

However, object adjectives have a very homogeneous definition and behaviour, while basic

[13] Tags in features follow the nomenclature in Table 5.13. The relevant tags are repeated here for clarity:

cd:	clause delimiter	aj:	adjective
dd:	definite determiner	av:	adverb
cn:	common noun	pe:	preposition
co:	coordinating element	ve:	verb
ey:	empty		

As explained in page 123 (example (5.3)), numbers indicate position counting from the adjective; a positive sign indicates position to the right of the adjective, a negative sign position to the left of the adjective.

adjectives exhibit more variability: some are gradable, some are not; some frequently occur in predicative position, some occur less frequently; etc. In addition, the basic class has both a positive definition (adjectives denoting an attribute from an ontological point of view) and a negative definition: all adjectives that do not fit in the constrained definition of object and event adjectives are assigned to the basic class. Therefore, it is not as homogeneous as the object class.

5.2.4.3 What about morphology?

Up to now, we have only used syntactic or, more generally, distributional features. Morphology is clearly related to the semantic classification proposed, as discussed in Section 3.6.2. One of the hypotheses of this thesis, as stated in Section 3.6.2, is that syntactic information is more reliable than morphological information for the semantic classification of adjectives. We therefore expect agreement between the clustering solution and the Gold Standard to be higher than the agreement with a classification based on morphological class.

An initial test with the present data seems to support this hypothesis. From the manual annotation in Sanromà (Sanromà, 2003), we mapped the classes as in Table 5.17, following the discussion in Section 3.6.2.

morph	**sem**
not derived	basic
denominal	object
deverbal	event
participle	event

Table 5.17: *Mapping from morphology to semantics.*

The agreement between this classification and the Gold Standard is $p_o = 0.65$ and $K = 0.49$. These figures are well beyond the baseline ($p_o = 0.49$ and $K = 10^{-15}$), but represent a lower accuracy than that obtained with distributional features ($p_o = 0.73$ and $K = 0.56$).

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 Gold Standard. Most of the mismatches are actually cases of polysemy (or ambiguity): for instance, *mecànic*, classified as basic in the Gold Standard, has a basic meaning (equivalent to ‘monotonous’) that was considered to be primary by the experts. However, it also has a ‘related to mechanics’ reading. 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. A principled investigation of the quantitative and qualitative differences between different levels of linguistic description, including syntax and morphology, is offered in Chapter 6.

5.3 Summary

The unsupervised experiments reported in this chapter had three main purposes: first, to provide feedback to the classification. The classification has been revised (one class was added and one class removed) according to the results of the initial clustering experiments. In the first experiment, a large overlap between the Gold Standard and the clustering solution has been observed, but by no means a one-to-one correspondence. In the second unsupervised experiment, with the modified classification, a clear correspondence between clusters and classes has been shown. Even if feature selection has been performed to select the most adequate features for our task, this piece of evidence provides support for the classification proposed.

The second goal was to gain insight into the modelling of polysemy. In experiment A, polysemy has been modeled in terms of additional classes. This approach is clearly not appropriate, at least for a clustering methodology: polysemous adjectives are not identified as a homogeneous and distinct class. Our explanation of these results is that polysemous adjectives are not a homogeneous and distinct class: they share behaviours of the classes in which they participate, but they do so to differing extents depending on their most frequent sense. In experiment B, only the main semantic class of each adjective has been explored, thus ignoring polysemy information. An alternative modeling of polysemy and design of the machine learning task should be envisaged. Chapter 6 examines one such alternative.

The third goal was to test the two models (self-defined semantic features and blind distributional features) of the data, so as to test their relative performance in the clustering experiments. The results should serve to provide an initial description of each class in terms of their distributional characteristics. The features were defined at a shallow level within a small window, with the reasoning that, adjectives exhibiting a limited syntactic variability, a small window would be enough for capturing their essential properties.

The examination of the whole set of distributional properties of adjectives (defined in terms of unigrams for experiment A, of pairs of bigrams for experiment B) was expected to provide additional clues with respect to the characterisation of each class. What we have found is that the predictions made from a theoretical point of view (semantic features in experiment A) are largely supported in clustering experiments using distributional features, and that parameters such as gradability, predicativity, and position with respect to the head are relevant no matter which representation of the data is chosen.

The analysis has also suggested new cues, such as the presence or absence of a determiner in the NP, the complementation patterns of adjectives, or the syntactic function of the head noun. These will be examined in Chapter 6.

In Experiment B, one class is not identified: event-related adjectives. The clustering only identifies event-related adjectives that are binary (bear complements). The remaining event adjectives seem to behave like basic ones. Event-related adjectives can not be characterised in terms of syntactic properties, as will be shown in Chapter 6.

In Chapter 4, we saw that the event class also presents problems from a semantic point of view. Thus, although some event-related adjectives present a distinct semantics with respect to basic and object classes, which was the reason to include them in the classification, they do not constitute a homogeneous class, neither from a semantic nor from a syntactic point of view.

Finally, the features examined within the clustering experiments are correlates of semantic and syntactic properties. However, the revised classification into basic, object-related and event-related adjectives bears an obvious relationship with morphological properties. We have performed an initial, simple test comparing the clustering solution for syntactic features and a naive mapping between morphology and semantics, based on the manual morphological classification by Sanromà (2003).

The results indicate that the morphology-semantics mapping achieves quite good results, but that better results are obtained with syntactico-semantic, distributional features. Our explanation is that the main source of mismatches between morphology and semantics are diachronic meaning shifts. If an adjective acquires a meaning that corresponds to another semantic class, it behaves like an adjective of the new class, irrespective of its morphological type. However, a systematic comparison of the roles of syntax, morphology and semantics for the classification of adjectives should be performed to test this hypothesis. Such a comparison is offered in Chapter 6.

Part of the material presented in this chapter has been published in the following articles:

Boleda, G. and Alonso, L. (2003). Clustering adjectives for class acquisition. In *Proceedings of the EACL'03 Student Session*, pages 9–16, Budapest, Hungary.

Boleda, G., Badia, T., and Batlle, E. (2004). Acquisition of semantic classes for adjectives from distributional evidence. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1119–1125, Geneva, Switzerland.

Chapter 6

Experiment C: Polysemy acquisition as multi-label classification

> ... high level semantic knowledge can be computed from large amounts of low-level knowledge (esentially plain text, part-of-speech rules, and optionally syntactic relations) ...
>
> Hatzivassiloglou and McKeown (1993, p. 172)

The experiments explained in Chapter 5 have served the purposes of refining the classification, providing an initial description of each class in terms of its distributional properties, and testing different models of the data: theoretically motivated features, n-gram, and bigram pair features. They also provide some support for the hypothesis that syntax is a better clue than morphology to the semantic class of adjectives. This hypothesis will be revised in this chapter.

In all the experiments explained so far, a pressing issue remains: polysemy. We have argued that treating polysemous classes as separate categories, as was done in Experiment A (Section 5.1) is not adequate to acquire data on polysemy.

This chapter reports on a series of experiments that set up an alternative architecture for the acquisition semantic classes for adjectives including polysemous class assignments. These experiments also aim at providing a thorough comparison of different linguistic levels of description (morphology, syntax, semantics) for the task at hand.

For these experiments, we use a supervised technique, Decision Trees, for two reasons. First, because the classification and our initial predictions with respect to the characteristics of each class have already been tested using an unsupervised technique, so that it is relatively safe to move to supervised techniques. Second, because it facilitates the new architecture to acquire information regarding polysemy.

Before turning to the discussion of the experiments, we explain and motivate the new architecture. Acquisition of semantic classes for polysemous words can be viewed as an instance of multi-label classification. Multi-label assignment has been tackled in recent years within the Machine Learning community mainly in the field of Text Categorisation, where a document can be described via more than one label, so that it effectively belongs to more than one of the targeted classes. Our situation is similar: polysemous adjectives are assigned to more than one class.

When discussing weight assignments for the weighted kappa measure in Section 4.3.6, we justified our weighting schema by modeling the decisions of the human judges. We argued that they could be viewed as independent decisions, because the task for an adjective was, for each pattern, to test whether the adjective fitted the pattern or not, irrespective of whether it fitted the previous or the following one.[1]

The same reasoning can be applied to the machine learning algorithm. Instead of attempting at a full-fledged classification (deciding, e.g., between B and BO), as in experiment A, the global decision can be decomposed into three binary decisions: Is it basic or not? Is it object-related or not? Is it event-related or not? The individual decisions can then be combined into an overall classification. If a lemma is classified both as basic and as object in each of the binary decisions, it is deemed polysemous (BO).

This approach has been the most popular one in Machine Learning when dealing with multi-label problems, according to, e.g., Schapire and Singer (2000) and Ghamrawi and McCallum (2005). It has recently been applied to other NLP problems, such as Semantic Role Labeling (Baldewein et al., 2004), entity extraction, and noun-phrase chunking (McDonald et al., 2005). McDonald et al. (2005) argue that multi-label classification is the most natural model to represent overlapping and non-contiguous segments in segmentation tasks. We will apply it here to the task of semantic class acquisition for adjectives: first, make a binary decision on each of the classes. Then, combine the classifications to achieve a final, multi-label classification.

This chapter is structured as follows. The classification and Gold Standard are explained in Section 6.1. The method followed for the experiments, from feature definition to experimental procedure, are detailed in Section 6.2. Section 6.3 reports the accuracy results and the error analysis, which are further discussed in Section 6.4.

6.1 Classification and Gold Standard

The classification pursued in this chapter is the same as in Experiment B (Section 5.2.1), namely, a classification into basic, object-related and event-related adjectives. However, we do not only aim at acquiring the main class of each adjective, as in Experiment B, but we also attempt at tackling polysemy.

The Gold Standard used in this chapter is the set of 210 adjective lemmata discussed in Chapter 4 (Gold Standard C in Table 4.25). The classification we rely on for the analysis of results is the one consensuated by the 3 experts.

Before moving to the experiment, we briefly review the characteristics of each semantic class and the predictions we can gather from Chapters 3 to 5.

We expect object adjectives to have a rigid position, right after a noun (in Catalan), with a strong adjacency constraint. Any other modifiers or complements (PPs, other adjectives, etc.) occur after the object adjective. This restriction also implies that object adjectives have very low frequencies for predicative positions. In addition, they are typically nongradable, or gradable only under very restricted circumstances (see examples in (3.44), page 56, and the subsequent discussion about this issue).

Event adjectives are the less clear class from a syntactic or distributional point of view. From the unsupervised experiments, it seems that they appear most naturally in predicative environments

[1] Remember that this is a simplified picture: as stated in footnote 14 (page 85), because a maximum of two answers could be given, the judges had to somehow “rank” the patterns and choose at most the two first patterns in the ranking. This means that the decisions are not entirely independent.

and tend to bear complements. This is probably due to the fact that most of them are deverbal and thus inherit part of the verbal argument structure. They tend to form larger constituents that are mostly placed in predicative position. For the same reason, they should appear in postnominal, rather than prenominal, position when acting as modifiers.

As for basic adjectives, most of them can be used nonrestrictively, although the most frequent modification is post-nominal. They are also used in predicative constructions. When combined with other kinds of adjectives, mainly object adjectives, they appear at the peripheria. Finally, they are typically gradable and thus co-occur with degree adverbs and suffixes.

6.2 Method

6.2.1 Linguistic levels of description and feature choice

In the experiments explained in Chapter 5, features from two levels of linguistic description were used: semantic (with the caveats mentioned in Section 5.1.2.1, page 105), and morphosyntactic or distributional features. In addition, a simulated classification was obtained based on morphology.

In this chapter, we test all these levels of description (plus an additional one) against each other. To that end, we use the same machine learning algorithm and feature selection mechanisms for all levels of description. Table 6.1 lists the linguistic levels considered: morphology (*morph*), syntax (*func*, *uni*, *bi*), and semantics (*sem*). In level *morph*, the features are categorical, that is, their values consist of a list of categories. In the remaining levels, the features are numeric (continuous), and are coded as proportions: their value corresponds to the proportion of contexts fulfilling the feature definition among all occurrences of a given adjective. Some exceptions to this definition will be signaled in the discussion that follows.

Level	Explanation	# Features
morph	morphological (derivational) properties	2
func	syntactic function	4
uni	uni-gram distribution	24
bi	bi-gram distribution	50
sem	distributional cues of semantic properties	18

Table 6.1: *Linguistic levels as feature sets.*

This section provides a detailed explanation of the feature definitions within each level. It also discusses the predictions made throughout this thesis against the empirical evidence gathered from analysing the distribution of the features in the Gold Standard. In the discussion, polysemous classes are separately considered for completeness. However, recall that in the machine learning experiments we will tackle each class (basic, event, object) as a separate learning task, so that we will not attempt at characterising polysemous classes in terms of homogeneous groups of adjectives.

6.2.1.1 Morphological features

Two morphological features are taken into account: derivational type (*dtype*) and *suffix*, in case the adjective is derived. The information is taken from the manually developed database by Sanromà (2003). The features and their possible values are listed in Table 6.2.

Dtype	Gloss	***Suffix*** **for dtype**
N	denominal	à, al, ar, ari, at, í, ià, ic, ista, ístic, iu, ós, ut, *other*
O	not derived	-
P	participial	t
V	deverbal	ble, er, iu, nt, or, ori, ós

Table 6.2: *Morphological features.*

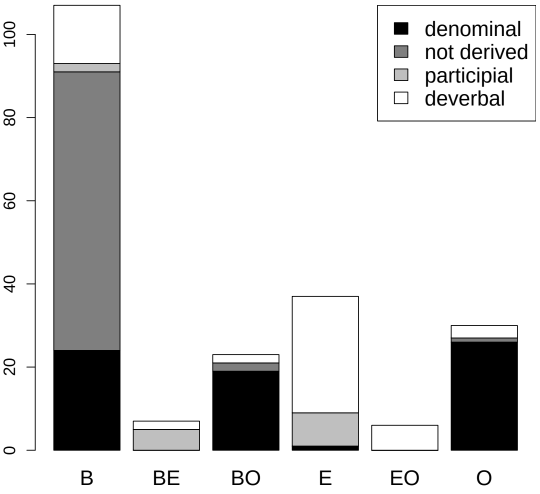


Figure 6.1: *Distribution of derivational type across classes.*

Note that feature *suffix* is a specialisation of feature *dtype* for the denominal and deverbal types; the question is what level of description (general type or particular suffix) will best serve the purposes of semantic classification.

Figure 6.1 shows the distribution of adjectives into semantic classes, according to their different derivational type. Table 6.3 contains the same information in form of contingency table (largest values boldfaced). The mapping predicted in Chapter 5 (not derived to basic, deverbal and participial to event, and denominal to object) receives support in Figure 6.1, although obvious mismatches also strike.

The best correspondence between morphology and semantic class is for not derived adjectives.

				class				
		B	BE	BO	E	EO	O	*Total*
dtype	denominal (N)	**24**	0	**19**	1	0	**26**	*70*
	not derived (O)	**67**	0	2	0	0	1	*70*
	participial (P)	2	5	0	**8**	0	0	*15*
	deverbal (V)	**14**	2	2	**28**	6	3	*55*
	Total	*107*	*7*	*23*	*37*	*6*	*30*	*210*

Table 6.3: *Distribution of derivational type across classes.*

	B	BE	BO	E	EO	O	Total
-	67	0	2	0	0	1	70
à	1	0	1	0	0	5	7
al	3	0	2	0	0	4	9
other	0	0	2	0	0	1	3
ar	6	0	1	0	0	2	9
ari	3	0	3	0	0	3	9
at	1	0	0	0	0	0	1
ble	3	1	0	6	1	0	11
er	0	1	0	1	0	1	3
í	0	0	0	0	0	1	1
ià	0	0	0	0	0	1	1
ic	1	0	2	0	0	6	9
ista	2	0	6	0	0	1	9
ístic	0	0	1	0	0	0	1
iu	3	0	2	1	4	2	12
nt	4	0	0	6	1	0	11
or	1	0	0	10	0	0	11
ori	2	0	0	2	0	1	5
ós	7	0	1	3	0	1	12
t	2	5	0	8	0	0	15
ut	1	0	0	0	0	0	1
Total	107	7	23	37	6	30	210

Table 6.4: *Distribution of suffix across classes.*

Only 3 out of 70 not derived adjectives are not purely basic. Deverbal and participial adjectives are mostly event-related. However, some (16) are only basic and some (13) are polysemous between event and another class. Note that most cases of polysemy between basic and event readings (BE) are due to participles. Section 6.2.1.2 further comments on this issue.

In contrast, all cases of polysemy between event and object readings (EO) are due to other kinds of deverbal adjectives. In fact, for deverbal adjectives there is some overlap with the object class (2 BO, 6 EO, 3 O adjectives), which does not happen with participials.

The worst correspondence affects denominal adjectives, as more adjectives are classified as basic (24 items) and BO (19 items) than as object (26 items). This is due to several factors: first, as we will shortly see, some denominal suffixes build mainly basic adjectives.

Second, recall that Raskin and Nirenburg (1998, p. 173) posit a productive semantic process that applies to all object-related adjectives so that they can be used as basic (see schema 3.45, page 57): from a "pertaining to X" meaning, a "characteristic of X" meaning is created, X being the object-denoting noun. Sometimes, the "pertaining to X" meaning is lost or nearly lost, so that only the basic meaning remains. This accounts, e.g., for *subsidiari* ('subsidiary'; does not mean 'related to subsidy', but 'secondary, complementary'). In most cases, however, the "pertaining to X" meaning remains, so that polysemy (or ambiguity) is created. For instance, *humà* ('human') is used in the study corpus both in the "pertaining to humans" sense (example (6.1a)) and the "characteristic of humans" (example (6.1b)) sense.

(6.1) a. [els animals domèstics] formen part de la societat humana
[the animals domestic] form part of the society human

‘Pets are part of the human society’

b. Qualsevol persona podia ser més intel·ligent que Swift, o almenys més humana
Any person could be more intelligent than Swift, or at-least more human

‘Anyone could be more intelligent than Swift, or at least show more humanity’

A particular case of this general semantic process are ideology-related adjectives, as has been discussed throughout the thesis.

Table 6.4 shows the distribution of adjectives across the semantic classes, according to the suffix they bear. Note that some suffixes correspond to the expected semantic class better than others, as has already been suggested. Within denominal adjectives, the suffixes that mostly map their adjectives to the object class are *-à*, *-al*, and *-ic*, and those that cause most mismatches are *-ar*, *-ista*, and *ós*. Within deverbal adjectives, *-or* is the most regular suffix, followed by the participial marker *-t*. Suffix *-iu* is the less regular one, with 3 basic, 2 basic-object, 4 event-object and 2 object adjectives; only one out of twelve deverbal adjectives with this suffix has been classified as purely event in the Gold Standard.

To sum up, although there is a quite regular relationship between morphology and semantic class, exceptions of various kinds arise for many adjectives. We expect semantic and syntactic information to improve on morphology for at least some of the semantic classes.

6.2.1.2 Syntactic features

Syntactic features come in three definitions. First, the syntactic function (level *func* in Table 6.1) of the adjective, as assigned by CatCG. Second, unigram distribution, as was defined in experiment A (Section 5.1): the parts of speech of the words preceding and following the adjective are separately encoded as features. Third, the bigram distribution. In Experiment B (Section 5.2.4), pairs of bigrams were explored, accounting for the left (3 tokens) and right (1 token) contexts. In that experiment, the bigrams corresponding to positions 2 and 3 to the left of the adjective were not as useful as the bigrams corresponding to the immediate context. Therefore, in the present experiments the first bigram is ignored: only the bigram around the target adjective is taken into account. In addition, only the most frequent 50 bigrams are considered, so as to avoid too sparse features.

For both unigrams and bigrams, the codification is as in experiment B: we use the tags in Table 5.13 (page 123), which encode more fine-grained information than the original part of speech tags. For discussion of the role of unigram and bigram features in the targeted classes, we refer the reader to Chapter 5. In this Section, we only discuss features regarding syntactic function in detail.

In the discussion of the syntactic and semantic features (this section and the next one), we will graphically show the distribution of the features across the classes using boxplots. The scale of all graphics is the same to facilitate comparisons between features, values ranging from 0 to 1 (with one exception that will be signaled). This kind of representation makes exploration easier than a numerical representation in a table. However, the mean and standard deviations of all features are available in Appendix B (Tables B.1 and B.2).

For all features discussed, we also perform one-way analysis of variance ANOVA, to check whether the difference in mean value of the feature across the 6 classes is significant. Homogeneity of variance is not assumed. Significant results (p < 0.05) provide evidence that the feature is relevant for at least one of the distinctions involved in our task. We also perform a two-tailed t-test on a specific distinction, basic vs. event, because, as we will see, basic and event adjectives are not often distinguishable on the basis of the syntactic and semantic features defined. Equality of variance is not assumed in these t-tests.

We now turn to the discussion of features regarding syntactic function. CatCG distinguishes among 4 syntactic functions for adjectives, as listed in Table 3.1 (page 20): post-nominal modifier, pre-nominal modifier, predicate in a copular sentence, and predicate in other environments. These will be used as features in the supervised experiments. Note that CatCG does not yield completely disambiguated assignment of syntactic functions (see Section 2.2.1). In case of ambiguity, we take the most frequent function among the alternatives left by CatCG, according to the frequencies in Table 3.1.

Figure 6.2 shows the distribution of feature values for each syntactic function (identified in the title of each graphic) in the different semantic classes. Table 6.5 reports the results of the ANOVA and t-tests for these features.

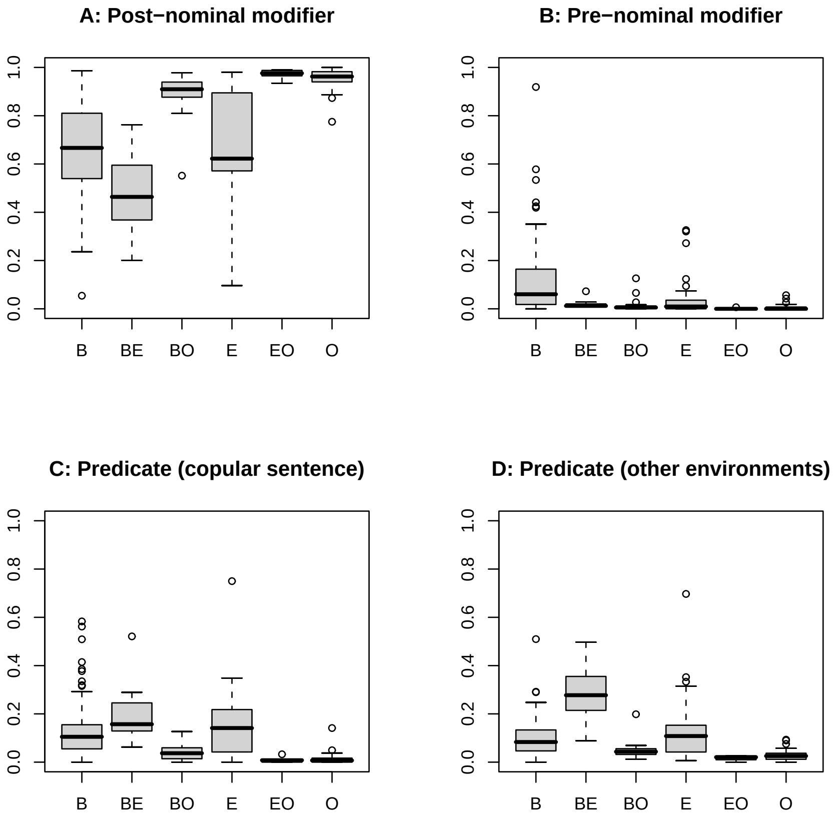


Figure 6.2: *Distribution of syntactic functions across classes.*

Object-related adjectives occur almost only as post-nominal modifiers (values close to 1 in graphic A for O, EO, and BO), and conversely have lower values for the remaining features. Only basic adjectives and some event adjectives occur as pre-nominal modifiers. Frequency

Comparison	Feature	df	F/t	p
6 classes	post-nominal mod.	5;36.5	64.4	$< 10^{-15}$
	pre-nominal mod.	5;42.3	15.8	$< 10^{-8}$
	pred (cop.)	5;35.8	27.3	$< 10^{-10}$
	pred (other)	5;36.1	25.0	$< 10^{-10}$
B, E	post-nominal mod.	54.6	-0.22	0.82
	pre-nominal mod.	108.8	3.56	0.0005
	pred (cop.)	52.8	-0.94	0.35
	pred (other)	44.5	-1.62	0.11

Table 6.5: *Results of statistical tests for syntactic function.*

of occurence as pre-nominal modifier, thus, can be used to some extent as a cue to distinguish basic and event adjectives, which is not the case with predicative functions. Note, however, that the feature is quite sparse, which favours confusion.

Both basic and event adjectives occur as predicates with a similar frequency (see graphics C and D), as confirmed by the t-tests in Table 6.5. Note, however, that adjectives classified as polysemous between basic and event (BE) have higher values for predicative functions than pure basic or pure events. The difference in mean values across classes B, BE, and E is only significant for predicative funcion in other environments (graphic D) according to one-way ANOVA: F(2;14.5)=7.33, p = 0.006. BE adjectives, thus, occur as adjunct predicates with a higher frequency than either basic or event adjectives.

Finally, note that the four features cover the whole distribution of adjective occurences in the corpus, so that one of them is necessarily redundant (its value can be deduced from the values of the other three features). However, a priori we cannot determine which one to eliminate, and it could be that different features were useful for the different distinctions to be made (basic or not, event or not, object or not). Therefore, we fed all features to the feature selection algorithm. The same applies to some of the features that will be presented in the next section.

6.2.1.3 Semantic features

The semantic features used for this chapter are a refinement and extension of the features used for Experiment A (see Section 5.1.2.1). They include the features defined for Experiment A at different levels of granularity, as well as additional features that have mainly arisen during the analysis of results of Experiments A and B. We next examine each of them in turn.

Not restrictive The cue to non restrictive use is again the pre-nominal modification tag assigned by CatCG. See the previous section for the distribution of this feature across classes.

Features regarding predicativity The cue for these features is also the function tag assigned by CatCG, except for two differences in granularity level. In Figure 6.2 (Section 6.2.1.2) it is clear that the overall distributions of the two predicative features are very similar, so that they convey similar information. Because both features are quite sparse, it makes sense to combine them. We define a feature *predicate (global)*, encoding the sum of the two predicative functions.

In Catalan, as in Spanish, two different copular verbs exist: *ser* and *estar*. Only some adjectives select for the copula *estar*. For Spanish, it has been argued that the distinction between the two roughly corresponds to whether the predicate denotes temporary or essential properties (Demonte, 1999, among many others). The behaviour of Catalan adjectives with respect to *ser* and *estar* is subject to much variability (De Cuyper, 2006), but is related to similar semantic information.

To check whether the selection of the copular verb correlates with an adjective's semantic class, predicative uses with *estar* as the main verb have been encoded as a separate feature. Note that CatCG does not provide dependency information. Therefore, a crude approximation has been used. For any adjective with a predicative function, a main verb is looked for (as tagged by CatCG) within seven words to the left of the adjective. Any of the verbs *estar*, *semblar* ('seem'), or *esdevenir* ('become') contribute to feature *estar* (*semblar* and *esdevenir* are similar to *estar* in terms of their aspectual properties).

Figure 6.3 shows the four features regarding predicativity that are included in the semantic level. Graphics A and B correspond to graphics C and D in Figure 6.2, and are included for comparison with *predicate (global)*. The global representation for predicativity (graphic C in Figure 6.3) yields sharper differences in the distribution of the feature across classes, and significant results are obtained in the 6-way comparison, as shown in Table 6.6. The same applies to feature *predicate (with* estar*)*, despite the fact that it is quite sparse. [2] However, neither of the two new features shows significant differences in the mean values for basic and event adjectives (see Table 6.6).

Comparison	**Feature**	**df**	F/t	**p**
6 classes	predicate (global)	5;36.8	44.62	$< 10^{-13}$
	predicate (with *estar*)	5;32.9	5.58	0.0008
B, E	predicate (global)	51.4	-1.59	0.12
	predicate (with *estar*)	43.2	-1.45	0.15

Table 6.6: *Results of statistical tests for features regarding predicativity.*

Features regarding gradability Three features to represent gradability have been explored here: *gradable* and *comparable*, defined as in Experiment A, and a feature *gradable (global)* which adds up the values for *gradable* and *comparable*. The reasoning is the same as for features regarding predicativity: features *gradable* and *comparable* convey similar information. The question for the experiments is which level of granularity is most adequate for each semantic class.

Figure 6.4 depicts the distribution of the three features regarding gradability across classes. Note that evidence for *gradable* and *comparable* is quite sparse, so that class distinctions for *gradable global* are clearer. However, the three features show highly significant mean differences according to the ANOVA test (Table 6.7). Object adjectives have clearly lower values for the three features than basic and event adjectives. If only basic and event adjectives are compared, the differences are not significant (see row *B, E* in Table 6.7), which means that basic and event adjectives do not differ in their mean gradability values.

[2] To conduct the ANOVA for feature *predicate (with* estar*)*, random noise was added, because adjectives in the O and EO classes all had a value of 0.

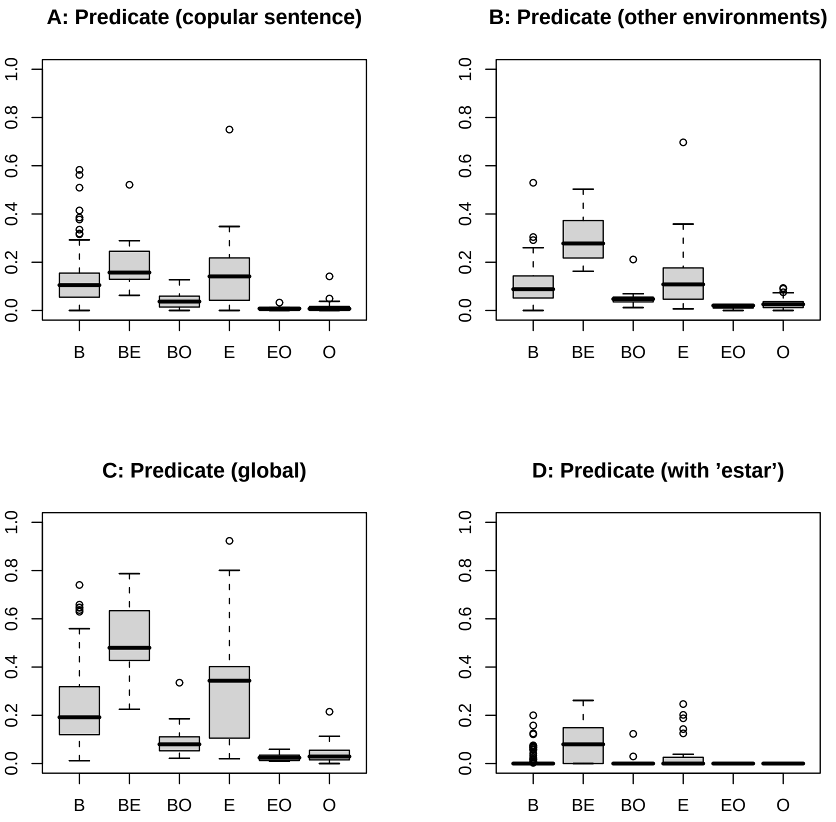


Figure 6.3: *Distribution of features regarding predicativity across classes.*

Features regarding definiteness In the analysis of Experiment A, the definiteness of the NP containing the adjective arose as an issue deserving closer inspection. We define three features to cover this aspect: *definite* (the adjective occurs within an NP headed by a definite determiner), *indefinite* (the adjective occurs within an NP headed by an indefinite determiner, or in a bare plural NP), and *bare* (the NP does not have a determiner). [3]

These features only make sense when adjectives occur within an NP, that is, when they functionally act as modifiers. To avoid unintended interactions with the dominant syntactic function of each adjective, which is separately accounted for in other features, the values for features regarding definiteness were computed as proportions within the total number of occurences of each adjective as a nominal modifier.

Because full-fledged phrase structure is not provided by CatCG, features regarding definiteness were implemented in terms of regular expressions searching for a determiner to the left of the adjective. Agreement properties (gender and number) of the determiner and the adjective were checked.

Figure 6.5 shows that object adjectives have clearly higher values for *definite* than basic and event adjectives (graphic A), which means that they appear in NPs with definite determiners

[3] As indefinite were considered the following lemmata (classified as *quantitative*, *indefinite*, and *cardinal* determiners in GLiCom's dictionary): *bastant*, *gaire*, *força*, *prou*, *massa*, *més*, *molt*, *tant*, *menys*, *poc*, *qualsevol*, *gens*, *tot*, *algun*, *cap*, *qualque*, *mant*, *sengles*, and the cardinals (*un* 'one', *dos* 'two', etc.).

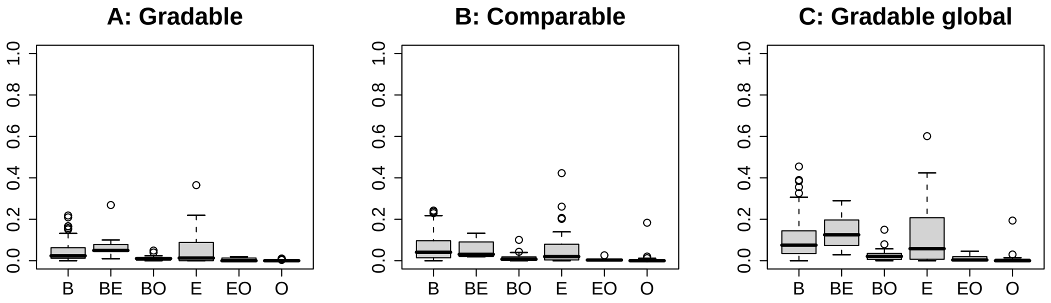


Figure 6.4: *Distribution of features regarding gradability across classes.*

Comparison	Feature	df	F/t	p
6 classes	gradable	5;26.6	21.54	$< 10^{-07}$
	comparable	5;35.9	14.06	$< 10^{-06}$
	gradable (global)	5;33.1	18.51	$< 10^{-07}$
B, E	gradable	44.8	-0.98	0.33
	comparable	49.5	0.41	0.69
	gradable (global)	48.5	-0.29	0.77

Table 6.7: *Results of statistical tests for features regarding gradability.*

with a much higher frequency than other kinds of adjectives. Conversely, they tend to have lower values for indefiniteness (graphic B), although the difference is not as sharp. Both features achieve highly significant results in ANOVA, as shown in Table 6.8.

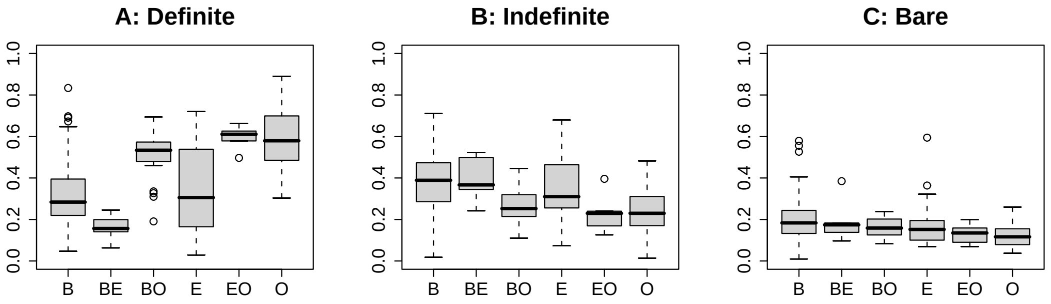


Figure 6.5: *Distribution of features regarding definiteness across classes.*

A possible explanation of these facts is as follows. Object adjectives include an object or object class in their denotation. They are frequently used to subclassify, that is, to specify subclasses (*subkinds*, according to McNally and Boleda (2004)) of the objects denoted by their head nouns. A definite determiner signals that the object is available in the context to both speaker and hearer, that is, it has already been introduced. Once a given class of objects is introduced in the discourse, a subclass is specified through the use of an NP containing an object adjective. Because the class has already been introduced, the new NP can be headed by a definite determiner.

Comparison	Feature	df	F/t	p
6 classes	definite	5;33.2	50.54	$< 10^{-13}$
	indefinite	5;28.4	9.34	$< 10^{-04}$
	bare	5;29	5.85	0.0007
B, E	definite	50.2	-0.70	0.49
	indefinite	58.6	1.29	0.20
	bare	64.8	1.58	0.12

Table 6.8: *Results of statistical tests for features regarding definiteness.*

A fictitious example of this phenomenon is given in (6.2). In (6.2a), the class of objects denoted by *malalties greus* ('serious illnesses') is introduced with a plural NP. Because pulmonary tuberculose is a subtype of serious illness, in (6.2b) the NP where *pulmonar* appears can be headed by a definite determiner. This explanation should be checked through an empirical study.

(6.2) a. Al congrés s'hi va parlar de malalties greus.
At-the conference REFL-CLIT PAUX-3ps speak about illnesses serious.

'At the conference, serious illnesses were talked about.'

b. Per exemple, la tuberculosi *pulmonar*.
For instance, the tuberculose pulmonary

'For instance, pulmonary tuberculose'

For feature *bare* (graphic C), the differences are lower than for the other two features. However, note that the median values depicted in the boxplots gradually descend. The differences for feature *bare* across all classes is significant (see Table 6.8), mainly due to the difference in mean between basic adjectives (0.20) and object adjectives (0.12). [4] Basic adjectives tend to appear in bare NPs with a slighty higher frequency than object adjectives.

Again, however, definiteness is not useful to distinguish between basic and event adjectives. The two classes do not exhibit significant differences in mean values for any of the features regarding definiteness, as shown in Table 6.8.

Syntactic function of the head noun In the analysis of Experiment A, the syntactic function of head nouns also emerged as a possible clue to the semantic class of adjectives. We define three features to account for this piece of information. As was the case with features regarding definiteness, only occurences of adjectives with a modifying function were taken into account, and the values were computed as proportions with respect to the total number of occurences with a modifying function.

CatCG only provides partial information for dependency: it only indicates whether the head noun is to be found at the left (post-nominal modifier) or the right (pre-nominal modifier) of the adjective. Therefore, we identified the head of the NP with some heuristics. The head was

[4] A t-test testing for the mean of feature *bare* for basic and object adjectives yields a highly significant result: t(87.9)=5.4, p $< 10^{-6}$.

searched for in the direction indicated by the function tag of the adjective. The first noun or determiner found was identified as the head of the adjective if it had compatible agreement (gender and number) features. If it did not, the search continued up to the limit of a 11-word window (7 to the left, 3 to the right of the adjective). With this procedure, we identified a head for 82% of the NPs in which our Gold Standard lemmata intervened. We did not assess accuracy.

Although the functional tagset of CatCG distinguishes between 10 possible functions for nominal categories, we only considere the three most frequent functions, namely, subject, object, and prepositional complement. *Prepositional complement* is used for nouns heading the NP complement of a prepositional phrase.

Figure 6.6 shows the distribution of the three features across classes. It is clear that the most useful feature is *head as prep. comp.*, for which object adjectives have much higher values than basic and event adjectives, as confirmed by ANOVA (Table 6.9). We have no ready explanation for the difference in values for feature *head as prep. comp.*. An examination of the structures involved, and most notably of the kinds of nominal heads, should be carried out to achieve further insight.

Feature *head as subject* yields mildly significant results (p=0.02), mostly due to differences between classes basic, event, and basic-event. Event adjectives modify subjects with a slightly higher frequency than basic adjectives (0.09 vs. 0.08). That this difference turns out to be significant in a t-test with the standard α=0.05 (p = 0.03; see Table 6.9) suggests that the proper α for this feature is lower. Note that this feature is quite sparse, and as a results variances are small.

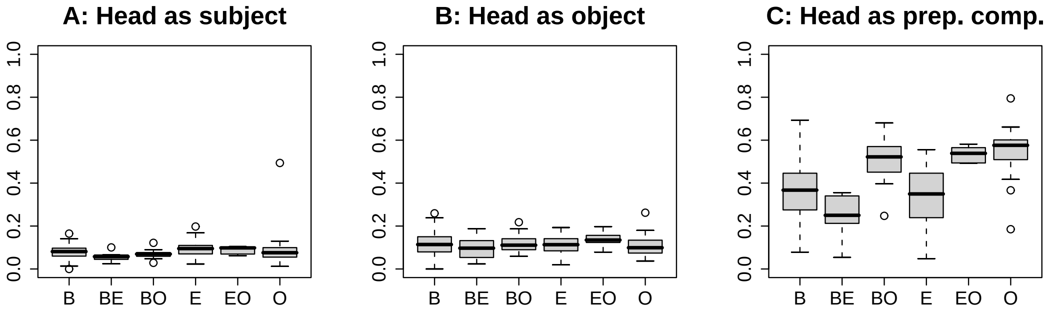


Figure 6.6: *Distribution of features regarding head syntactic function across classes.*

Comparison	Feature	df	F/t	p
6 classes	head as subject	5;28.7	3.42	0.02
	head as object	5;27.5	0.68	0.64
	head as prep. comp.	5;31.7	30.11	$< 10^{-10}$
B, E	head as subject	50.9	-2.18	0.03
	head as object	70.8	0.34	0.73
	head as prep. comp.	65.3	1.26	0.21

Table 6.9: *Results of statistical tests for features regarding head syntactic function.*

Distance to the head noun In this experiment, we attempt at generalising feature *adjacent* from Experiment A. That feature identified adjectives occuring between a noun and another adjective (see Section 5.1.2.1). The hypothesis is that object adjectives occur more tightly attached to the noun than other types of adjectives. Therefore, in the new definition we counted total distance to the head noun, expecting it to be lower for object adjectives than for other classes. The head of the NP was identified using the heuristics explained above, and the distance was counted in terms of number of words. Therefore, the values of this feature do not range between 0 and 1, as the other features, but between 0 (immediately adjacent) to 1.83 (almost two words between adjective and head).

As can be seen in Figure 6.7 and Table 6.10, this feature effectively distinguishes between object and either basic or event, but does not help distinguish basic from event.

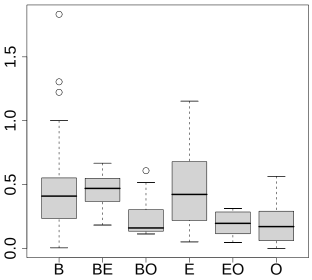


Figure 6.7: *Distribution of feature* distance to head noun *across classes.*

Comparison	Feature	df	F/t	p
6 classes	distance to head noun	5;30.6	14.58	$< 10^{-06}$
B, E	distance to head noun	56.7	-0.89	0.38

Table 6.10: *Results of statistical tests for feature* distance to head noun.

Binaryhood In Experiment B (Section 5.2), the “binaryhood” of an adjective was revealed as a clue to its belonging to the event class. Event adjectives tend to bear complements, a correlate of their being binary (having two arguments), while basic and object adjectives most frequently have only one argument.

Because most adjective complements are headed by a preposition, this feature corresponds to the proportion of prepositions found at the right of a given adjective. Clausal complements are introduced by a conjunction instead, but conjunctions were not included in the feature definition because they proved too noisy: in many cases, a conjunction following an adjective does not introduce a complement, but an independent clause (as in constructions like *és clar que* . . . , ‘it is clear that . . . ’).

As shown in Figure 6.8 and Table 6.11, feature *binary* has a different distribution when the overall, 6-way classification is considered, and also when only basic and event adjectives are considered.

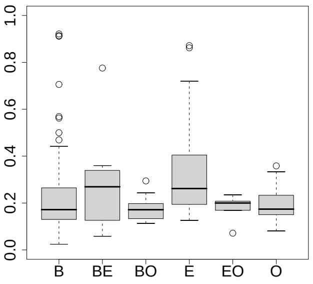


Figure 6.8: *Distribution of feature* binary *across classes.*

Comparison	Feature	df	F/t	p
6 classes	binary	5;29.1	5.24	0.001
B, E	binary	54.6	-3.09	0.003

Table 6.11: *Results of statistical tests for feature* binary.

Gender and number Finally, we tested the distribution of agreement features, because it could be that semantic properties encoded in terms of number and gender would correlate with the semantic class of an adjective. We only code plural and feminine, but note that singular and masculine convey exactly the same information as their counterparts. Some adjectives are underspecified with respect to gender. For these cases, we used a default value estimated from the frequencies of all adjectives in the corpus, namely, 0.48 for feminine.

Figure 6.9 and Table 6.12 show that only one tendency is observed, namely, for object adjectives to modify feminine heads in a higher proportion than the rest of the adjectives (note the mild significance of ANOVA for feature *feminine*).

This could be due to the fact that in Catalan, abstract nouns denoting concepts such as science disciplines are typically feminine. As has been mentioned above, object adjectives are typically used to identify subclasses, and thus are used with these kinds of nouns. This suggests that the semantic type of the head is relevant for the semantic classification of adjectives, an issue that deserves further research.

However, agreement features are clearly not as useful as the other features examined for the distinctions we target at.

Comparison	Feature	df	F/t	p
6 classes	plural	5;26.9	1.08	0.39
	feminine	5;27.3	2.76	0.04
B, E	plural	58.3	-1.38	0.17
	feminine	77.6	-0.44	0.66

Table 6.12: *Results of statistical tests for features regarding agreement.*

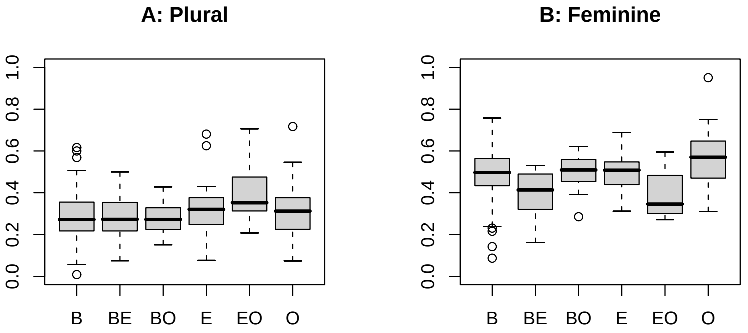


Figure 6.9: *Distribution of features regarding agreement across classes.*

6.2.1.4 Discussion

We have discussed 19 different features from the syntactic function (*func*) and semantic (*sem*) levels of description. [5] Only 3 of the 19 features, namely, *pre-nominal modification/not restrictive*, *head as subject*, and *binary*, exhibit significant mean differences (measured at the 0.05 level) between the basic and the event class, as summarised in Table 6.13. In fact, we have expressed doubts about the usefulness of feature *head as subject*, because the differences in mean value are very small.

Feature	**df**	F/t	**p**
pre-nominal mod./not restrictive	108.8	3.56	0.0005
head as subject	50.9	-2.18	0.03
binary	54.6	-3.09	0.003

Table 6.13: *Syntactic and semantic features to distinguish basic and event adjectives.*

This leaves us with only two generalisations about the difference between basic and event adjectives, namely, that basic adjectives appear in pre-nominal modification with a higher frequency than event adjectives, and that event adjectives bear complements in a larger proportion of cases than basic adjectives. There could be a relationship between these two properties: if event adjectives enter into heavier constructions, bearing complements themselves, they are most easily placed after the noun. The two features are negatively correlated (Pearson's ρ=-0.24, t(208)=-3.57, p< 0.0005), which provides support for this explanation.

However, the differences between the two classes are very weak: there are numerous exceptions to the two generalisations provided, as can be gathered from Figures 6.2 and 6.8, and the remaining features do not help distinguish between these classes. In contrast, 17 out of the 19 features discussed exhibit significant mean differences for at least one of the distinctions in terms of semantic classes, and most notably serve to distinguish object adjectives from basic and event.

We can expect the accuracy of experiments ran on the basis of syntactic or semantic features to yield much worse results for the basic or event classes than for the object class. This prediction

[5] There are 4 different syntactic features and 18 semantic features; however, 3 of them overlap.

is confirmed in the results of the experiments (see Section 6.3).

Finally, note that for almost all features, polysemous adjectives in which the object class intervenes (BO and EO) pattern together with object adjectives, while BE adjectives pattern together with basic and event adjectives. Also, note that BE adjectives have a distinct behaviour for some features: they exhibit higher values than either basic or event for features regarding predicativity, lower values for feature *indefinite*, and lower values for *head as prep. comp.* and *feminine* (see Figures 6.5, 6.6, and 6.9).

The seven BE adjectives in the Gold Standard are: *animat* ('animated | cheerful'), *cridaner* ('vociferous | loud-coloured'), *embolicat* ('wrapped | embroiled'), *encantat* ('charmed | absent-minded'), *obert* ('opened | open'), *raonable* ('reasonable'), *sabut* ('known | wise'). These are mainly participials derived from stative verbs. In contrast, participles that are classified as pure events in the Gold Standard tend to be derived from dynamic verbs: *acompanyat* ('accompanied'), *encarregat* ('entrusted'), *orientat* ('oriented'), *picat* ('perforated, minced, crossed'), *promès* ('promised'), *recomanat* ('recommended'). Only verbs *oblidar* (for participle *oblidat*, 'forgotten') and *irar* (for participle *irat*, 'angry') are stative, in the sense that they cannot be used in the progressive.

In Chapter 4, we argued that the Aktionsart of the verb plays a role in the semantic type of the deriving adjectives. The data reviewed in this Section seems to support this hypothesis, although further research is needed to clarify the relationship between the two factors.

6.2.2 Machine Learning algorithm: C4.5

The technique we use for the present supervised experiments is one of the most widely used machine learning techniques, namely, Decision Trees (Witten and Frank, 2005). We carry out the experiments with the Weka software package (see Section 2.2.3).

Decision Trees provide a clear, compact representation of the data at hand, and thus facilitate inspection of results and error analysis. The particular algorithm chosen, Weka's J48, is the latest open source version of Ross Quinlan's C4.5 (Quinlan, 1993). C4.5 is a very popular algorithm "which, with its commercial successor C5.0, has emerged as the industry workhorse for off-the-shelf machine learning" according to Witten and Frank (2005, p. 189). It has been used in related research, most notably in the already cited study by Merlo and Stevenson regarding the classification of English verbs into unergative, unaccusative and object-drop (Merlo and Stevenson, 2001).

The algorithm works in a top-down fashion. It recursively selects an attribute to split up the examples in groups that are as pure as possible in terms of class, until all objects below a particular leave of the tree are of the same class, or no further splits are possible.

The functioning is clearest with nominal features, such as our morphological features (*dtype* and *suffix*). In an imaginary decision tree, the algorithm first splits on *dtype*, so that each of the possible values (denominal, not derived, deverbal, and participial) constitutes a new branch. Within each branch, the class of the adjective is directly determined (e.g., for branch *not derived* the algorithm would stipulate class basic, because no further information is available) or further branches are created with the *suffix* feature.

For instance, as has been shown in Section 6.2.1.1, it makes sense within the denominal type to classify adjectives bearing suffixes *-ar* and *-ós* as basic. To do so, a branch would be created for each possible value of feature *suffix*, class basic would be assigned to leaves *-ar* and *-ós*, and class object to the remaining denominal suffixes.

The measure of purity that is used to select the best feature to branch on is based in the Information Theory measure of entropy that has been discussed in Section 4.5.2 (see Equation (4.13), page 96). In this case, entropy measures the uncertainty of the class of the objects below a particular node in the tree. The less homogeneous the classes, the higher the uncertainty and thus the entropy value, and the lower the purity of that node.

To select the feature with highest purity (lowest entropy), the *information gain* measure is used. The information gain computes the difference between the average entropy of the tree prior to branching and the average entropy of the tree after branching on a particular feature. The feature that provides the highest information gain is selected to branch on. This criterion, however, favours features with a large number of values. To correct for this situation, the *gain ratio* is used. The gain ratio takes into account the number and size of the daughter nodes so as to penalise features with a large number of values.

This method only works with nominal features. To extend it to numeric features, only binary splits are considered. Objects are ordered according to their numeric values, and breakpoints are considered that do not separate objects of the same class. The information gain is computed in the usual way for each breakpoint, and a threshold is set so that objects with a value higher than the threshold are separated from objects with a value lower than the threshold.

Let's assume, for instance, that all basic adjectives have values higher than 0.1 for feature *pre-nominal modifier*, and the remaining adjectives have values lower than 0.1. A decision tree built with this feature could separate basic adjectives from the rest by establishing a threshold of 0.1 on feature *pre-nominal modifier*. Because different thresholds can identify different distinctions, numeric features can be used more than once in a particular tree (nominal features are only used once).

Trees that are built with this procedure tend to overfit the training data they are based on. Therefore, a *pruning* strategy is usually designed. In C4.5, as in most decision tree applications, the pruning occurs after the tree has been built (an alternative is to pre-prune, or stop developing a particular subtree when it starts overfitting the data). The pruning occurs with two operations: *subtree replacement* and *subtree raising*. In subtree replacement, a particular subtree is replaced by a single leave, proceeding from leaves to root. In subtree raising, a whole subtree A replaces another subtree B. The objects originally under subtree B are re-classified in subtree A. Because it is a time-consuming operation, it is used with some restrictions (see Witten and Frank (2005, p. 193), for details).

To decide whether a particular pruning operation is worth performing, C4.5 estimates the error rate that would be caused by a particular node if an independent test set were considered. However, the estimate is based on the training data. Witten and Frank (2005, p. 194) state that the method "is a heuristic based on some statistical reasoning, but the statistical underpinning is rather weak and ad hoc. However, it seems to work well in practice." For details on this method and more details on decision tree induction and C4.5, see Quinlan (1993) and Witten and Frank (2005, Sections 4.3 and 6.1).

Because the focus of our research is the comparison between levels of linguistic description, we did not perform parameter optimisation. The parameters for the machine learning experiments were set with the default options of C4.5 as implemented in Weka. These are listed in Table 6.14. The remaining options available in Weka for the C4.5 (J48) algorithm are not used in the

Parameter	Gloss	Value
confidenceFactor	confidence factor used for pruning	0.25
minNumObj	minimum number of instances per leaf	2
subtreeRaising	whether to consider subtree raising when pruning	TRUE

Table 6.14: *Parameters for J48/C4.5.*

default case.

6.2.3 Feature selection

Decision trees select the best feature to split on at each branch of the tree. It seems, thus, that the algorithm itself performs feature selection, so that irrelevant features will not be used in a given tree. However, as the tree is built, less and less objects are available for the algorithm to select further features to split the data on. At some point, an irrelevant attribute will be chosen because it happens, just by chance, to divide the objects in an adequate way. This point is inevitably reached: note that the minimum number of objects per leaf is set to 2 (see Table 6.14), so that if, say, 4 or 5 objects are found below a node, the algorithm will still build a further branch to better classify them. Despite the pruning procedures, some amount of error remains.

Irrelevant attributes typically decrease performance by 5 to 10% when using Decision Trees (Witten and Frank, 2005, p. 288). Therefore, it is advisable to perform feature selection prior to actual machine learning experiments, so that only the most relevant features are passed on to the machine learning algorithm.

Witten and Frank (2005) state that “the best way to select relevant attributes is manually, based on a deep understanding of the learning problem and what the [features] actually mean”. This supports the definition of carefully designed feature sets, as has been done for semantic features in Experiments A and C. Further automatic feature selection helps reduce dimensionality and potentially increases accuracy even for these feature sets, and is necesary for n-gram features.

There are several possible approaches to feature selection (Witten and Frank, 2005, Section 7.1, offer an overview). After several trials of the different implementations available in Weka, we chose a method (called *WrapperSubsetEval* in Weka) that selects a subset of the features according to the performance of the machine learning algorithm using only that subset. Accuracy for a given subset of features is estimated with cross validation over the training data. Because the number of subsets increases exponentially with the number of features, and because n-fold cross validation involves n actual experiments for each possible subset, this method is computationally very expensive.

We use a common search strategy to lessen the computation time involved in feature selection. The search strategy starts with no attributes, and adds one attribute at a time (*forward selection*). The accuracy measured in terms of cross validation is used to check whether a given attribute improves on the accuracy of the subset previously evaluated. Accuracy has to increase by at least a pre-specified threshold (0.01 in our case) to allow a feature to be added. In typical forward selection, if no feature improves accuracy, the search terminates. This procedure finds a local, but not necessarily a global, maximum. We use a more sophisticated variant, which keeps track of a number of subsets evaluated, and revisits the next subset with highest performance instead of terminating the search when performance starts to drop. This method is called *best-*

first search (*BestFirst* in Weka).

The parameters involved in the feature selection algorithm and the values chosen (which correspond to the default values in Weka) are depicted in Table 6.15. The value n for parameter *lookupCacheSize* stands for the total number of attributes in each data set.

Parameter	Gloss	Value
folds	number of cross validation folds	5
seed	seed used to randomly generate cross validation splits	1
threshold	minimal improvement in accuracy to select feature	0.01
lookupCacheSize	maximum size of the lookup cache of evaluated subsets	n
searchTermination	maximum number of subsets to re-evaluate before terminating search	5

Table 6.15: *Parameters for feature selection.*

6.2.4 Testing for difference across linguistic levels

One of the aims of the PhD is to test the strengths and weaknesses of each level of linguistic description for the task at hand. Within our setting, the most natural way to carry out this test is to compare the accuracy obtained with each of the feature sets in the machine learning experiments.

Accuracy scores are obtained through comparison with the Gold Standard classification. It is equivalent to the proportion of observed agreement (p_o) measure discussed in Section 4.3.1, but it is customarily reported as a percentage, rather than a proportion.

The comparison seems quite straightforward to carry out: run an experiment with the features corresponding to each of the linguistic levels, and see which ones get better results. However, to be sure that the results are not particular to the setup of the experiment (particularly, to the choice of training and testing subset), a significance test should be used.

One of the most popular choices to perform accuracy comparisons in the field of Machine Learning was for some time a paired t-test using the outcomes of several random partitions into test and train sets. In Machine Learning, the aim is usually to test differences between algorithms, not feature sets. With this method, for algorithm A, n different accuracy scores are obtained based on n different partitions of the data set. The same n partitions are used to obtain accuracy scores for algorithm B. A standard paired t-test is applied on the accuracy scores to test whether the difference in mean values is significant or not.

Dietterich (1998) showed that this kind of test has an elevated Type I error probability, that is, it predicts that there is a significant difference when there is not in a higher percentage of cases than that established in the confidence level. Indeed, increasing the number of partitions eventually leads to a significant result in most cases, and there is no principled way to determine the optimal number of partitions to use.

The source of the elevated Type I error is the violation of the independence assumption underlying the t-test. The accuracy scores obtained through repeated holdout (several random partitions into train and test sets) are not independently obtained, because the same data is repeatedly used either in the train or the test sets. This results in an underestimation of the variability both in training and testing data. Using cross validation instead of repeated holdout smooths the violation (at least the test data is not reused), but the training data from one fold to another is shared to an important extent (80% in the usual 10-fold choice), which also implies a violation of the independence assumption.

A number of approaches that have lower Type I error have been proposed. An influential proposal was recently made by Nadeau and Bengio (2003). They alter the t-test formula so that increasing the number of partitions does not result in a higher significance. The standard paired t-test formula is as follows:

$$t = \frac{\bar{d}}{\sqrt{\frac{\sigma_d^2}{k}}} \tag{6.1}$$

$\bar{d}$ is the average of the differences observed in two paired samples (for instance, reaction time before and after taking a pill for a number of patients), σ_d^2 the variance of the differences and k the degrees of freedom (number of differences - 1). In our case, $\bar{d}$ is the mean difference between the accuracies obtained with one algorithm and the accuracies obtained with another algorithm on the same partitions of the dataset, σ_d^2 the variance of these differences and k corresponds to the number of partitions used (less one).

The correction by Nadeau and Bengio (2003) involves a change in the denominator (correcting the variance estimation) so that the t statistic cannot be increased simply by increasing the value of k:

$$t = \frac{\bar{d}}{\sqrt{(\frac{1}{k} + \frac{n_2}{n_1})\sigma_d^2}} \tag{6.2}$$

In Equation (6.2), n_1 is the proportion of instances that are used for training and n_2 the proportion that are used for testing. Nadeau and Bengio (2003) advise using this test, named *corrected resampled t-test*, with 15 random train-test partitions. Witten and Frank (2005), and the associated Weka software package, adopt this kind of test but use it with 10-run, 10-fold cross validation (see Section 6.2.6 for an explanation of this term), which they say “is just a special case of repeated holdout in which the individual test sets for *one* cross-validation do not overlap” (Witten and Frank, 2005, p. 157). This is the implementation we will use here, for which k = 99, n_2/n_1 = 0.1/0.9, and $\bar{d}$ and σ_d^2 are obtained from 100 differences, one for each fold.

As Witten and Frank (2005) observe, this approach is just a heuristic that has been shown to work well in practice. No procedure to significance testing with data obtained from a single dataset has been devised that is theoretically sound. For a recent proposal that alters the sampling scheme instead of correcting the formula of the t-test, see Bouckaert (2004).

Note, however, that the corrected resampled t-test can only compare accuracies obtained under two conditions (algorithms or, as is our case, feature sets). We test more than two linguistic levels, so that ANOVA should be performed, instead of a t-test. However, standard ANOVA is not adequate, for the same reasons for which the standard t-test is not adequate. In the field of Machine Learning, there is no established correction for ANOVA for the purposes of testing differences in accuracy (Bouckaert, 2004; Witten and Frank, 2005). Therefore, we will use multiple t-tests instead. This increases the overall error probability of the results for the significance tests.

Finally, to provide a basis for all the comparisons, we provide a baseline accuracy to test the machine learning results against. We will use the same baseline as in Experiment B, namely, to assign all lemmata to the most frequent class (the mode). Weka provides an algorithm that performs just this classification, termed *ZeroR* (for "zero rule"), which we use to obtain the baseline accuracy estimates.

6.2.5 Combining features of different levels

The levels of linguistic description used for the experiments explained in this section will be separately tested. However, it is natural to think that a combination of all features should improve upon the results obtained with each of the levels. We have performed one experiment to test the degree of improvement when combining features from different levels. We thus add the feature set *all* to the five individual feature sets listed in Table 6.1.

Recall that feature selection is performed for each of the experiments. There are in total 95 features for the whole experiment, and not all of them are equally useful for each of the distinctions. To test how much improvement is obtained using all levels simultaneously, we only use the features that are selected by the feature selection algorithm at least 30% of the time, that is, in at least 30 out of the 100 experiments performed for each class.

The features selected vary according to the targeted class. They are listed in Table 6.16. In the table, each feature is followed by the absolute number of times it is selected for the supervised experiments. Below each class name, the total number of features used for the class in level *all* is depicted. [6]

Class	Level	Features
B	morph	*dtype* (99)
5	func	*pre-nominal mod.* (83)
	uni	*-1av* (47), *+1ve* (42)
	bi	*-1ve+1co* (76)
	sem	*not restrictive* (87)
E	morph	*dtype* (97)
7	func	*post-nominal mod.* (80)
	uni	-
	bi	*-1co+1co* (44), *-1cd+1pe* (40)
	sem	*pred (global)* (43), *pred (with* estar*)* (40), *gradable* (32)
O	morph	*suffix* (89)
19	func	*post-nominal mod.* (87), *pred (other)* (45), *pred (cop.)* (42)
	uni	*+1pe* (65), *-1ve* (62), *-1cn* (55), *-1pe* (43), *+1ve* (36)
	bi	*-1cn+1ve* (92), *-1cn+1aj* (69)
	sem	*pred (cop.)* (76), *binary* (66), *pred (other)* (56), *pred (global)* (55), *bare* (51), *not restrictive* (46), *definite* (45), *head as prep. comp.* (44), *pred (with* estar*)* (43), *gradable (global)* (30)

Table 6.16: *Most frequently used features.*

[6] Note that in some cases the same feature is selected in different levels: features *pre-nominal mod.* and *not restrictive* have exactly the same definition, and features *pred (cop.)* and *pred (other)* are used in the *func* and *sem* levels. To perform the experiments, one of them is randomly discarded. The number of features depicted below each class name does not take into account repetitions.

This table gives a kind of definition of each of the classes in terms of the features used. Note that for the *morph* level, the derivational type is the right level of description for the basic and event classes, but not for the object class, which needs more fine-grained information, namely, the suffix (further discussion in Section 6.3.2).

As for contextual features, they quite closely match the predictions made so far, as well as the results of the unsupervised experiments reported in Chapter 5. For basic adjectives, the most informative features are the position with respect to the head noun (in levels *func* and *sem*), gradability (level *uni*: *-1av* encodes contexts in which an adverb precedes the adjective), and predicativity (level *bi*: *-1ve+1co* translates as 'adjective preceded by verb and followed by coordinating element'). Unigram feature *+1ve* ('followed by verb') seems to point to noun heads acting as subjects, but, as discussed in Section 6.2.1, this feature does not successfully characterise basic adjectives. Other interpretations of this feature remain open.

As expected, event adjectives are much less homogeneous with respect to contextual features: the most frequently selected features are selected in less than 50% of the folds in levels *func*, *bi*, and *sem*, and no feature reaches the 30% threshold for level *uni*. Surprisingly enough, neither *binary* nor *not restrictive* are often chosen to characterise this class. However, bigram *-1cd+1pe* ('preceded by clause delimiter and followed by preposition') points to binaryhood of the adjective, and also to the use of event adjectives in absolute constructions. Both features are typical of participial adjectives. Information about predicativity and gradability is used to identify event adjectives in level *sem*.

As could be expected from the discussion, for object adjectives contextual information is most useful, as shown by the fact that 18 contextual features are used over 30% of the time for tree construction, as opposed to only 4 and 6 for basic and event, respectively. The most used piece of evidence, as in all experiments performed so far, is the post-nominal position. In level *func*, the relevant feature is *post-nominal mod.*. In levels *uni* and *bi*, it is considered in features containing *-1cn* ('preceded by common noun').

Some of the remaining pieces of information used for object adjectives coincide with the "classical" properties discussed in the literature: predicativity (*pred* and *-1ve* features), gradability (*gradable (global)*), and adjacency constraints (*-1cn1aj*, 'preceded by common noun and followed by adjective'). Note that feature *distance to the head noun* is not used.

However, for this class it is also useful to consider pieces of information that are not often tackled in theoretical literature, such as the definiteness of the NP where the adjective occurs (features *bare*, *definite*), presence (absence) of prepositional complements (*+1pe*, *binary*), or function of the head noun (*head as prep. comp.*). The unigram *+1ve* feature is also frequently selected, as it was for class basic.

A final note on methodology: the actual experiments with the *all* level were performed with the same procedure as the experiments with the rest of the levels, namely, using feature selection prior to building the decision trees. This way, the accuracies for each level of description are obtained under the same circumstances.

6.2.6 Summary

The whole experimental procedure is as follows. We carry out experiments using 6 different levels of linguistic description (*morph*, *func*, *uni*, *bi*, *sem*, and the combination of the 5, *all*), and compare them to a "zero rule" baseline that uniformly assigns the most frequent class to all adjectives.

To perform the experiments, we randomly partition the data into 10 different groups of the same size (21 adjectives per group). In a particular experiment, 9 groups (189 adjectives) are used for feature selection and training, and 1 group is used for testing (*test set*). Only the features selected by the feature selection algorithm explained in Section 6.2.3 are used to build the tree. 10 experiments are performed, using a different group as test set each time. This procedure is commonly referred to as 10-fold cross validation.

The whole procedure is repeated 10 times (10 *runs*), with different random partitions, so that in the end 100 experiments are performed for each of the six feature sets plus the baseline. This method is called 10-run, 10-fold cross validation (*10x10 cv* for short). Because a separate set of experiments is carried out for each of the three classes (basic, event, object), 2100 different experiments are performed: 10 runs x 10 folds x 7 levels x 3 classes. The significance of the differences in mean accuracy is tested using the corrected resampled t-test explained in Section 6.2.4.

Although we perform a separate experiment for each class (basic or not, event or not, object or not), as has been explained in the introduction to this chapter, in the end we want a combined classification. We achieve a full classification by merging the decisions on the individual classes. Suppose that for a given adjective the decisions of the algorithm are yes for basic, no for event, and no for object. In merging the decisions, we end up with a monosemous assignment to basic for that adjective. If the decisions are yes for basic, no for event, and yes for object, we classify the adjective as polysemous between basic and object (BO).

If all three decisions are negative for a particular adjective (so that no class is provided), we assign it to the most frequent class, namely basic. If they are all positive, we randomly discard one, because class basic-event-object is not foreseen in our classification. Less than 5% of the merged class assignments obtained in the experiment are a no-class (679 out of 14700) or basic-event-object (41 out of 14700) assignment.

For each adjective, however, only 10 different class proposals are obtained for each linguistic level, because each adjective is only used once per run for testing. The accuracy of the different linguistic levels for full classification is assessed comparing 10 accuracies. In this case, in the formula (6.2) (page (6.2)), k = 9, n_2/n_1 = 0.1/0.9, and $\bar{d}$ and σ_d^2 are obtained from 10 differences, one for each run.

6.3 Results

It took over 68 hours (almost 3 days) for the 2100 experiments to run on a PC with one Pentium 4 processor (with 2.4GHz CPU and 524MB of memory), and a Fedora 2.6 Linux operative system. Table 6.17 shows the number of features[7] and the time (in hours and minutes) that it took for each level to complete the 10x10 cross validation experiment.

The nomenclature in Table 6.17 will be used throughout the analysis of results. Level *bl* corresponds to the baseline, *all* to the combination of all features, and the remaining levels follow the nomenclature in Table 6.1 (page 133).

[7] For the *all* level (combination of all features), the number of features is the mean of the number of features for each class, shown in Table 6.16.

Level	#Feat	Hours	Mins.
bl	0	0	5
morph	2	0	18
func	4	0	57
uni	24	14	16
bi	50	40	20
sem	14	7	26
all	10.3	5	19

Table 6.17: *Running time needed for each linguistic level.*

Due to the feature selection procedure, the time increases almost linearly with the number of features; for the morphological level, with only 2 features, the time needed to complete the 300 experiments (100 per class) is 18 minutes. For the bigram level, with 50 features, the experiments last over 40 hours. Recall that, as explained in Section 6.2.3, the feature selection algorithm performs 5-fold cross validation for each subset of features evaluated, and that the number of subsets increases exponentially with the number of features. The forward selection, best-first search strategy used for feature selection allows the corresponding increase in time to be linear rather than exponential. Nevertheless, the method used is computationally very intensive and only feasible for small feature sets, such as the ones we use.

6.3.1 Accuracy results

The accuracy results for each of the binary decisions (basic or not, event or not, object or not) are depicted in Table 6.18. Each column contains the mean and standard deviation (signaled with ± and in smaller font) of the accuracy for the relevant level of information over the 100 results obtained with 10x10 cv.

	Basic	Event	Object
bl	65.2±11.1	76.2±9.9	71.9±9.6
morph	72.5±7.9	89.1±6.0	84.2±7.5
func	73.6±9.3	76.0±9.3	81.7±7.4
uni	66.1±9.4	75.1±10.6	82.2±7.5
bi	67.4±10.6	72.3±10.2	83.0±8.3
sem	72.8±9.0	73.8±9.6	82.3±8.0
all	**75.3**±7.6	**89.4**±5.7	**85.4**±8.7

Table 6.18: *Accuracy results for binary decisions.*

The table shows that the baseline is quite high: assigning all adjectives to the most frequent class yields a mean accuracy of 65.2 for class basic, 76.2 for event, and 71.9 for object. Note that for the binary decisions, the baseline assignment is *basic* (for classification in basic or not), *not event* (for class event) and *not object* (for class object). Only a quarter of the adjectives are event-related adjectives, so that a uniform assignment to class *not event* gets three quarters of the job done.

As could be expected, the best results are obtained with the *all* level (bold faced in Table 6.18), which is a combination of features from the rest of the levels. This level achieves a mean improvement of 12.3% over the baseline. The greatest improvement (13.5%) is obtained for class object; the lowest improvement (10.1%) is obtained for class basic.

The differences in accuracy results between most levels of information are, however, rather small. Table 6.19 shows two-by-two comparisons of the accuracy scores obtained with each level. Each cell contains the mean and the level of significance of the differences in accuracy between a given level (row) and another level (column). The significance is marked as follows: * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$. If no asterisk is shown, the difference is not significant.

class	**level**	**bl**	**morph**	**func**	**uni**	**bi**	**sem**
basic	morph	7.2					
	func	8.3	1.1				
	uni	0.9	-6.4	-7.5			
	bi	2.1	-5.1	-6.2	1.3		
	sem	7.6	0.3	-0.8	6.7	5.4	
	all	10.1	2.9	1.8	9.2	8	2.5
event	morph	13*					
	func	-0.1	-13.1*				
	uni	-1.1	-14*	-1			
	bi	-3.9	-16.9**	-3.8	-2.8		
	sem	-2.4	-15.4**	-2.3	-1.3	1.5	
	all	13.2*	0.3	13.4*	14.3*	17.1**	15.7**
object	morph	12.3*					
	func	9.8	-2.5				
	uni	10.3	-2	0.5			
	bi	11.1*	-1.2	1.3	0.9		
	sem	10.4*	-1.9	0.6	0.1	-0.7	
	all	13.5*	1.1	3.7	3.2	2.3	3

Table 6.19: *Binary decisions: Comparison of the accuracies per linguistic level.*

The difference between each of the levels and the baseline is shown in the first column in the table. For class basic, although all levels seem to improve over the baseline, no improvement is significant according to the corrected resampled t-test. For the event class, only levels *morph* and *all* offer a significant improvement in accuracy over the baseline; the remaining levels even obtain a slightly lower accuracy score. Finally, for class *object*, all levels except for *func* and *uni* achieve a significant improvement over the baseline.

The discussion in Section 6.2.1 has shown that contextual features are most accurate for the object class, and that they in general do not help distinguish between basic and event adjectives. It is to be expected that the best results with contextual features are obtained with the object class.

In Section 6.2.1 we have also shown that the morphology-semantics mapping works best for event adjectives, and worst for object adjectives. As for basic adjectives, although almost all non derived adjectives are basic, quite a large number of denominal and deverbal adjectives are also basic, so that the *morph* level does not achieve such positive results as with the other two classes.

As for comparisons between levels, the only significant differences concern the event class: levels *morph* and *all* are significantly better than the remaining levels, and comparably so (the difference between *all* and *morph*, 0.3%, is not significant). For neither of the other two classes (basic and object) are significant differences in accuracy observed when using different levels of information.

To sum up, the best raw results are obtained with the *all* level. Our feature sets improve upon the baseline for two of the classes (event and object). For the event class, only levels *morph* and *all* improve upon the baseline; they do so by about 13%. For the object class, all levels except for *func* and *uni* are significantly better than the baseline, and all levels show a comparable improvement of around 11%.

These results concern the individual decisions. However, in the end we do not want three separate decisions, but a single classification including polysemy information. Table 6.20 shows the accuracy results for the classification obtained by merging the three individual decisions for each adjective. The resulting classification for each level of description for the lemmata in Gold Standard C are provided in Appendix C (Section C.3.2).

Recall from Section 6.2.4 that accuracy is the percentage version of p_o. Given the nature of our task, it makes sense to compute weighted versions of accuracy. We report three accuracy measures, full, partial, and overlapping. They correspond to the different agreement definitions used for interrater agreement (p_o, wp_o, op_o), as explained in Chapter 4. Full accuracy requires the class assignments to be identical. Partial accuracy gives some credit to partial agreement, following the weighting scheme in 4.14 (page 85). Overlapping accuracy only requires some overlap in the classification of the machine learning algorithm and the Gold Standard for a given class assignment to count as correct.

In Chapter 4, we have argued that for our task, the most appropriate measure of agreement is partial agreement. It can also be argued that partial accuracy is the most appropriate measure for evaluation of machine learning algorithms in our task. A class assignment that presents some overlap with the Gold Standard (even if it is not identical to the Gold Standard assignment) is more useful than a class assignment with no overlap for any purpose the classification is used. Full accuracy does not account for this aspect.

However, an overlapping class assignment is not as useful as an identical class assignment. Overlapping accuracy does not distinguish between the two cases. In Chapter 4, we have shown that weighted measures of agreement (wp_o and wK) using the weighting scheme in Table 4.14 achieve a balance between the strictest and most relaxed extremes represented by full and overlapping definitions of agreement. Thus, the estimate for agreement (here, for accuracy) is more realistic. The reasoning behind the weighting scheme can be applied to the machine learning task, as in our architecture the three classes are independently acquired.

In Table 6.20, the baseline has dropped, because the difficulty of the task has increased. The most frequent class, basic, accounts for only half of the items. The baseline increases in each of the three conditions (full, partial, overlapping), as the constraints on the definition of accuracy relax. Therefore, in each of the conditions it is harder to obtain significant improvements.

Again, the best results are obtained with the *all* level. The second best results are obtained with level *morph* in all three conditions. This result could be expected from the results obtained in the individual decisions (Tables 6.18 and 6.19); however, note that the differences between levels are much clearer in the merged classification than in binary decisions.

	Full	**Partial**	**Overlapping**
bl	51.0±0.0	60.5±0.0	65.2±0.0
morph	60.6±1.3	78.4±0.5	87.8±0.4
func	53.5±1.8	70.9±1.0	79.8±1.3
uni	52.3±1.7	68.2±0.8	76.7±1.0
bi	52.9±1.9	68.3±1.3	76.9±1.8
sem	52.0±1.3	69.6±1.1	78.7±1.7
all	**62.3**±2.3	**80.9**±1.5	**90.7**±1.6

Table 6.20: *Accuracy results for merged classification.*

Table 6.21, containing the average differences between levels and their level of significance, confirms the analysis just made. Under the strictest evaluation condition (full accuracy), only levels *morph*, *func*, and *all* significantly improve upon the baseline. Levels *morph* and *all* are better than the remaining levels, to a similar extent (*all* achieves on average 1.7% more accuracy than *morph*, but the difference is not significant).

agreement	**level**	**bl**	**morph**	**func**	**uni**	**bi**	**sem**
full	morph	9.7***					
	func	2.5*	-7.1***				
	uni	1.4	-8.3***	-1.1			
	bi	2	-7.7***	-0.6	0.6		
	sem	1	-8.7***	-1.5	-0.4	1	
	all	11.4***	1.7	8.9***	10***	9.4***	10.4***
partial	morph	17.9***					
	func	10.4***	-7.5***				
	uni	7.7***	-10.2***	-2.7**			
	bi	7.8***	-10.1***	-2.6**	0.1		
	sem	9.1***	-8.8***	-1.3	1.3	1.2	
	all	20.4***	2.5*	10***	12.7***	12.6***	11.3***
overlapping	morph	-22.6***					
	func	14.6***	-8***				
	uni	11.4***	-11.1***	-3.1**			
	bi	11.7***	-10.9***	-2.9**	0.2		
	sem	13.4***	-9.1***	-1.1	2.0	1.8	
	all	25.4***	2.9*	10.9***	14***	13.8***	12***

Table 6.21: *Full classification: Comparison of the accuracies per linguistic level.*

In the partial and overlapping evaluation conditions, all levels achieve a highly significant improvement over the baseline ($p < 0.001$). Therefore, the classifications obtained with any of the levels are more useful than the baseline, in the sense that they present more overlappings with the Gold Standard.

Levels *morph* and *all* are better than all the other levels, and level *all* improves upon than *morph*. Level *func* is better than *uni* and *bi*. That is, the 4 features corresponding to the main syntactic functions of the adjective achieve better results than the 24 and 50 n-gram features. However, level *sem*, consisting of carefully handcrafted features, does not improve on any of the other levels, and performs consistently worse than *morph* and *all*. We believe this is due to the difficulties in distinguishing basic and event adjectives, as will be discussed in the next section.

Figure 6.10 graphically shows the differences between levels reported in Tables 6.20 and 6.21. Each line corresponds to the accuracies obtained with each level in the 10 runs of the experiment. In graphic A (full accuracy), only levels *all* and *morph* are clearly above the baseline, and they do not exhibit clear differences in accuracy. In graphics B (partial accuracy) and C (overlapping accuracy), there are three clear groups with respect to performance: the baseline, well above it the contextual levels (*func*, *uni*, *bi*, *sem*), and well above them levels *morph* and *all*. Level *all* is clearly better than *morph* under the latter two evaluation conditions.

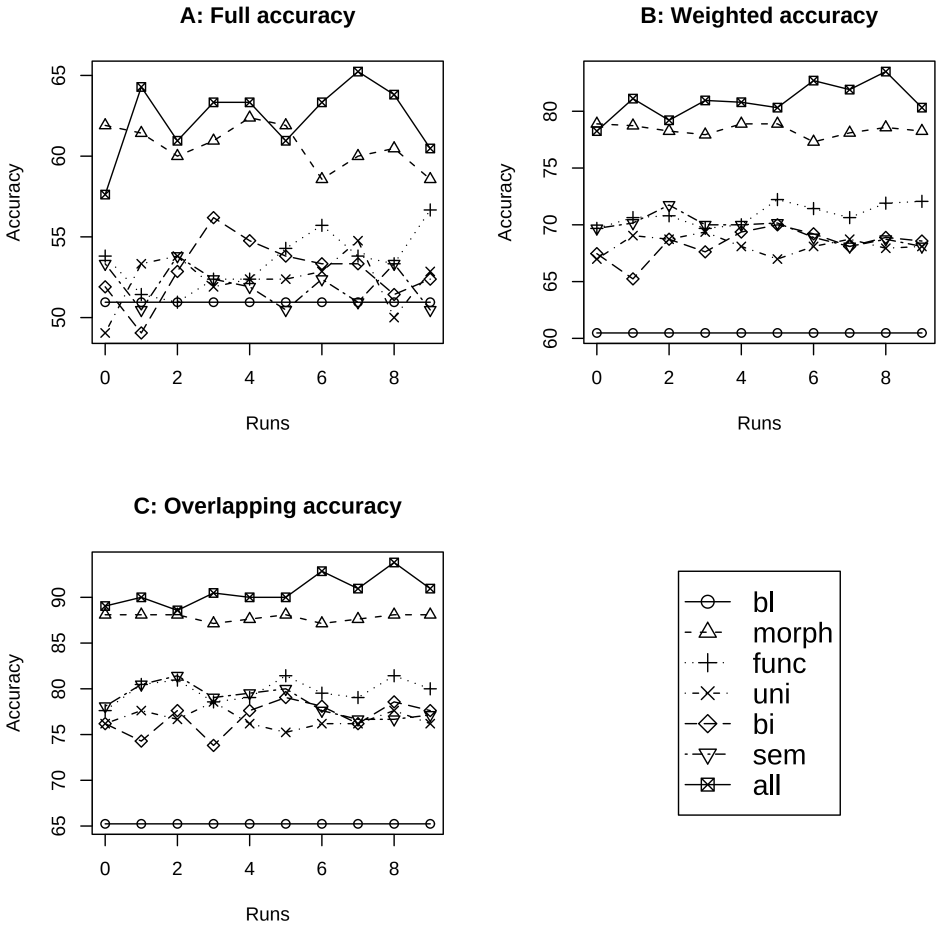


Figure 6.10: *Full classification: Comparison of the accuracies per linguistic level.*

To sum up, if we take partial accuracy as a reference for evaluation, all levels of analysis improve upon the baseline in the task of semantically classifying adjectives. However, levels including morphological information are clearly superior to levels using only contextual or distributional information: levels *morph* and *all* achieve a mean improvement of 19% over the baseline; using levels *func, uni, bi,* and *sem*, the improvement is halved (average 8.8%).

The best result obtained for the full classification of adjectives with our methodology achieves a mean 62.5 (full accuracy) or 80.9 (partial accuracy), which represents an improvement of 11.4% and 20.4% over the baseline, respectively. To our knowledge, no sufficiently similar tasks have been attempted at in the literature for comparison. We report the results of two tasks that share different aspects with our own enterprise.

First, in the SensEval-3 competition, a task for Catalan was organised (Màrquez et al., 2004) involving the disambiguation of 5 Catalan adjectives with an average of 4 senses each: *natural* (‘natural’), *popular* (‘popular’), *simple* (‘simple’), and *verd* (‘green’). The corpora used to train the systems were obtained from news corpora, and the less frequent senses of the words to be tagged were discarded. The baseline used, 71.7% f-score, was obtained by assigning the most frequent sense to all occurrences. The best system achieved an f-score of 86.5%, which represents an improvement of 14.8 over the baseline. This task concerns Catalan adjectives. However, within a word disambiguation system, our research would concern the first step, namely, defining the inventory of senses, not the actual disambiguation, which is the aspect discussed in Màrquez et al. (2004).

A more similar piece of research is reported in the already cited study by Merlo and Stevenson (2001), which discusses the automatic classification of English verbs into unergative, unaccusative and object-drop. The part of speech and the language are different. The number of core classes is the same, but they do not model polysemy information, so that their system performs 3-way, not 6-way classification. In addition, the training material is balanced (20 verbs in each class), which results in a lower baseline (34%). However, the task also involves acquisition of lexical knowledge, and they use the same machine learning algorithm, namely C4.5. Their system achieves 69.8% accuracy, which given the differences explained is comparable to our 62.5-80.9%.

6.3.2 Error analysis

The two best feature sets are clearly *morph* and *all*. Therefore, the error analysis that follows focuses on these two levels.

As mentioned in Section 6.2.2, decision trees readily lend themselves to exploration. For error analysis, we first examine the decisions made by the algorithm. We report decision trees built on all training data, because they can be viewed as the best trees the algorithm can build, as they use all evidence available for the tree construction. The decision trees built for evaluation only use 90% of the data, and it is difficult to establish a criterion to single one out (among the 100 trees built for each level and class) for examination. It should be borne in mind that although the trees reported in what follows are representative of the kinds of trees built in each level, the trees built for evaluation were subject to variability.

Figure 6.11 shows the decision trees built for the basic and event classes. In the decision trees shown in this figure, as in the remaining decision trees depicted in this chapter, B stands for basic, E for event, O for object, and the same letters preceded by letter *n* stand for not basic, not event, and not object.

basic	**event**
`dtype = N: B (70.0/27.0)`	`dtype = N: nE (70.0/1.0)`
`dtype = O: B (70.0/1.0)`	`dtype = O: nE (70.0)`
`dtype = V: nB (55.0/18.0)`	`dtype = V: E (55.0/19.0)`
`dtype = P: nB (15.0/7.0)`	`dtype = P: E (15.0/2.0)`

Figure 6.11: *Level* morph*: Decision trees for classes basic and event.*

Note that the two trees in Figure 6.11 are symmetric: the tree for basic states that denominal (N) and not derived (O) adjectives are basic, while deverbal (V) and participial adjectives (P) are not. The tree for event states just the reverse: denominal and not derived map to not event, and deverbal and participial to event. The tree for event corresponds to the expectations. Note, however, that the leave that accumulates most mistakes is the one that maps deverbal adjectives to the event class: 19 out of 55 deverbal adjectives are not event adjectives.

As for the tree corresponding to the basic class, we would expect denominal adjectives to map to not basic, and indeed this is the node that causes most mistakes (27 out of 70 items under this leave are not basic). However, due to the productive semantic process explained in Section 3.6 (schema 3.45 in page 57), which creates a basic adjective from an object adjective, as well as to the cases of ambiguity caused by nationality-type adjectives, many cases of polysemy between basic and object exist. Also, as explained in Section 6.2.1.1, some suffixes seem to build mainly basic adjectives. As a result, the best generalisation for denominal adjectives when tackling the basic/not basic distinction is a mapping to basic.

The fact that some suffixes build mainly basic adjectives also causes the tree for the object class, shown in Figure 6.12, to use suffix information, not derivational type as for the other two classes. Suffixes forming denominal adjectives are mapped to the object class (as boldfaced in Figure 6.12) except for suffixes *ar* and *ós* (in italics). Suffixes forming deverbal adjectives, as well as absence of suffix (for not derived adjectives), are mapped to not object, as expected.

```
suffix = -:  nO (70.0/3.0)
suffix = à:  O (7.0/1.0)
suffix = al:  O (9.0/3.0)
suffix = other:  O (3.0)
suffix = ar:  nO (9.0/3.0)
suffix = ari:  O (9.0/3.0)
suffix = at:  nO (1.0)
suffix = ble:  nO (11.0/1.0)
suffix = er:  nO (3.0/1.0)
suffix = í:  O (1.0)
suffix = ià:  O (1.0)
suffix = ic:  O (9.0/1.0)
suffix = ista:  O (9.0/2.0)
suffix = ístic:  O (1.0)
suffix = iu:  O (12.0/4.0)
suffix = nt:  nO (11.0/1.0)
suffix = or:  nO (11.0)
suffix = ori:  nO (5.0/1.0)
suffix = ós:  nO (12.0/2.0)
suffix = t:  nO (15.0)
suffix = ut:  nO (1.0)
```

Figure 6.12: *Level* morph*: Decision tree for class object.*

The data support the analysis made in the previous Section. We again see that the most compact, less error-prone representation in terms of morphological features corresponds to the event class. The decision tree for the basic class is also based on the derivational type, but the mismatches are more frequent. The object class needs a finer-grained level of information, namely suffix, because no uniform mapping is obtained for the derivational type. Also note that the trees depicted in Figures 6.11 and 6.12 are flat: they have only one level of branching.

We next examine the decision trees for the *all* level, which are more complex than the trees for the *morph* level. The trees for classes basic and event, depicted in Figures 6.13 and 6.14, also branch first on the derivational type. However, within each derivational type, contextual restrictions are made to account for distributional cues indicating a shift from the expected class.

In class basic (Figure 6.13), the refinement concerns the three primary properties of basic adjectives: predicativity, gradability, and nonrestrictive use. Note that denominal adjectives that are neither predicative nor gradable are deemed not basic with high accuracy (line 3), while those that are predicative and gradable are deemed basic also with high accuracy (line 9).

Similarly, deverbal adjectives that are not gradable are deemed not basic, and those that are gradable are deemed basic (lines 12 and 13). The latter decision causes quite a high error (17 out of 35 adjectives under that leave are not basic), which shows that quite a large number of event adjectives are gradable. Finally, participial adjectives that can be used nonrestrictively tend to be basic (line 17), while those that cannot tend not to be basic (line 15).

```
 1 dtype = N
 2 |   -1ve+1co <= 0.001527
 3 |   |   -1av <= 0.101517: nB (17.0/1.0)
 4 |   |   -1av > 0.101517: B (2.0)
 5 |   -1ve+1co > 0.001527
 6 |   |   -1av <= 0.026168
 7 |   |   |   -1ve+1co <= 0.012658: nB (7.0/1.0)
 8 |   |   |   -1ve+1co > 0.012658: B (5.0/1.0)
 9 |   |   -1av > 0.026168: B (39.0/4.0)
10 dtype = O: B (70.0/1.0)
11 dtype = V
12 |   -1av <= 0.054288: nB (20.0)
13 |   -1av > 0.054288: B (35.0/17.0)
14 dtype = P
15 |   not restrictive <= 0.001022: nB (5.0)
16 |   not restrictive > 0.001022
17 |   |   not restrictive <= 0.016839: B (5.0)
18 |   |   not restrictive > 0.016839
19 |   |   |   -1ve+1co <= 0.03629: B (2.0)
20 |   |   |   -1ve+1co > 0.03629: nB (3.0)
```

Figure 6.13: *Level* all*: Decision tree for class basic.*

The decision tree for class event (Figure 6.14) with level *morph* is the best of the three trees, so that little additions in terms of contextual cues are made in the *all* level, as shown in Figure 6.14. As could be expected, the refinement concerns the most error-prone leave, namely, the mapping from deverbal to event. The refinement uses information on binaryhood (presence of a preposition to the right).

We would expect binary adjectives to be classified as event. However, this piece of evidence is not robust, as has been discussed in Section 6.2.1, so that the constraints are quite strange, establishing a value either below 0.00058 or above 0.003 to indicate eventhood (lines 5 and 7). This does not look like a promising generalisation. In fact, we have argued that no contextual cues can identify the event class in its present definition. The mean improvement of the *all* level over the *morph* level for the event class is negligible (0.3%; see Table 6.21).

```
1 dtype = N: nE (70.0/1.0)
2 dtype = O: nE (70.0)
3 dtype = V
4 |   -1cd+1pe <= 0.003053
5 |   |   -1cd+1pe <= 0.000581: E (38.0/13.0)
6 |   |   -1cd+1pe > 0.000581: nE (5.0)
7 |   -1cd+1pe > 0.003053: E (12.0/1.0)
8 dtype = P: E (15.0/2.0)
```

Figure 6.14: *Level* all*: Decision tree for class event.*

Finally, the tree for the object class is the only tree to branch first on a non-morphological feature, namely, the syntactic feature *post-nominal modifier*. It establishes a generalisation consistent with the theoretical characterisation of the class, namely, that the adjectives that do not act as post-nominal modifiers in a high proportion of cases (almost 86% of the occurrences) are not object (line 1). Within the remaining adjectives, the tree establishes a distinction in terms of suffix (same as in Figure 6.12), and an additional constraint according to the binaryhood of an adjective (lines 8 and 9), which does not make sense from a theoretical point of view, and indeed seems to be quite *ad hoc*, given the small number of lemmata concerned.

```
 1 post-nominal mod. <= 0.859155: nO (124.0/3.0)
 2 post-nominal mod. > 0.859155
 3 |   suffix = -: nO (14.0/2.0)
 4 |   suffix = à: O (6.0)
 5 |   suffix = al: O (7.0/1.0)
 6 |   suffix = other: O (3.0)
 7 |   suffix = ar
 8 |   |   -1pe <= 0: nO (4.0/1.0)
 9 |   |   -1pe > 0: O (2.0)
10 |   suffix = ari: O (8.0/2.0)
11 |   suffix = at: O (0.0)
12 |   suffix = ble: O (1.0)
13 |   suffix = er: O (1.0)
14 |   suffix = í: O (1.0)
15 |   suffix = ià: O (1.0)
16 |   suffix = ic: O (8.0)
17 |   suffix = ista: O (6.0)
18 |   suffix = ístic: O (1.0)
19 |   suffix = iu: O (8.0/1.0)
20 |   suffix = nt: nO (3.0/1.0)
21 |   suffix = or: nO (8.0)
22 |   suffix = ori: O (1.0)
23 |   suffix = ós: O (2.0)
24 |   suffix = t: nO (1.0)
25 |   suffix = ut: O (0.0)
```

Figure 6.15: *Level* all*: Decision tree for class object.*

A second source of insight is the comparison of the errors made by the two sets of features. We have built 10 different classifications (corresponding to each of the runs) for each of the linguistic levels used. To achieve a unique classification for comparison with the Gold Standard, we apply majority voting across all runs.

The accuracies for the classifications obtained via majority voting are shown in Table 6.22, which includes the accuracies for the 10x10 cv experiment for comparison (data as in Table 6.20, page 158). For both levels *morph* and *all*, in italics in Table 6.22, results obtained with majority voting are worse than those obtained through 10x10 cv. However, accuracies decrease to a greater extent for the *all* level: 5.2% in the full evaluation condition, as opposed to only 1.6% for level *morph*. As a result, in the classification obtained through majority voting, level *morph* performs equally or slightly outperforms level *all*, depending on the evaluation scheme chosen.

	Full		**Partial**		**Overlapping**	
	10x10 cv	**Majority**	**10x10 cv**	**Majority**	**10x10 cv**	**Majority**
bl	51.0±0.0	51.0	60.5±0.0	60.5	65.2±0.0	65.2
morph	*60.6*±1.3	*59.0*	*78.4*±0.5	*77.5*	*87.8*±0.4	*86.7*
func	53.5±1.8	55.2	70.9±1.0	69.5	79.8±1.3	76.7
uni	52.3±1.7	52.4	68.2±0.8	67.3	76.7±1.0	75.2
bi	52.9±1.9	52.4	68.3±1.3	66.2	76.9±1.8	73.8
sem	52.0±1.3	53.8	69.6±1.1	65.4	78.7±1.7	71.4
all	*62.3*±2.3	*57.1*	*80.9*±1.5	*76.3*	*90.7*±1.6	*86.7*

Table 6.22: *Accuracy results for merged classification (majority voting).*

The data in Table 6.22 explain the fact that in the majority voting classification, level *all* makes a slightly higher number of mistakes (89) than level *morph* (86).

Table 6.23 compares the Gold Standard classification (rows) with the *all* and *morph* classifications (columns). Diagonal elements (matches) are in italics, and off-diagonal cells representing the largest numbers of mismatches are boldfaced.

		all						**morph**					
		B	BE	BO	E	EO	O	B	BE	BO	E	EO	*Total*
GS	B	*94*	**12**	0	0	1	0	*82*	2	**10**	**11**	2	*107*
	BE	1	*6*	0	0	0	0	0	*1*	0	6	0	*7*
	BO	**16**	1	*5*	1	0	0	5	0	*16*	2	0	*23*
	E	5	**23**	1	*7*	1	0	4	7	0	*25*	1	*37*
	EO	0	2	0	0	*4*	0	0	0	0	6	*0*	*6*
	O	**16**	1	6	2	0	*5*	6	0	**21**	3	0	*30*
	Total	*132*	*45*	*12*	*10*	*6*	*5*	*97*	*10*	*47*	*53*	*3*	*210*

Table 6.23: *Levels* all *and* morph *against Gold Standard.*

As can be seen in Table 6.23, both classifications fare quite well with the basic class. Level *morph* does better than level *all* in classes BO and event. The *all* level does better than *morph* in polysemous clases BE and EO, and slightly better in class object.

In fact, level *morph* does not produce a single object-only classification, due to the fact that it maps most denominal adjectives to both basic and object (see trees in Figures 6.11 and 6.12). As a result, the BO class is overly large in the *morph* level (47 items, as opposed to 23 in the Gold Standard). 16 out of the 23 BO lemmata are correctly classified by level *morph*. However, because of the overgeneration of BOs, 31 lemmata that are tagged as either basic or object in the Gold Standard are also assigned to BO.

In contrast, level *all* is overly discriminative: most of the BO cases (16 out of 23) are assigned to basic by the *all* level. These involve most ideology-related adjectives, such as *anarquista*, *comunista*, *feminista* (‘anarchist(ic)’, ‘communist(ic)’, ‘feminist(ic)’, ‘socialist(ic)’), as well as adjectives *diürn* (‘diurnal’) and *nocturn* (‘nocturnal’). As discussed in Chapters 3 and 4, these adjectives are controversial, in that they seem to involve ambiguity rather than true polysemy. However, also typically polysemous adjectives are classified as basic: *familiar*, *humà*, *infantil* (‘familiar | of family’, ‘sensitive | human’, ‘childish | infantile’).

Similarly, level *all* shows difficulties in spotting object adjectives (16), which it classifies as basic. These involve some controversial cases, such as *barceloní*, *manresà*, *renaixentista* (‘of Barcelona/Manresa’, ‘related to the Renaissance’), which can be said to pattern with nationality-denoting and ideology adjectives. Also, *fangós* (‘full of mud’), discussed in Section 4.5, and *diari*, *estacional* (‘daily, seasonal’), which involve a notion of frequency and not only a pure reference to an object. It seems reasonable that these adjectives share some distributional properties with basic objects. However, seemingly uncontroversial object adjectives are also involved, such as *epistemològic*, *mercantil*, and *ontològic* (‘epistemological’, ‘mercantile’, ‘ontological’).

The migration of BO and object adjectives to basic in the *all* level shows that contextual properties tend to “inflate” the basic class to a size greater than it already has (132 adjectives in the *all* level, 107 in the Gold Standard). In Chapter 5, we argued that the basic class is the default class, in the sense that adjectives that do not fit into the narrower definitions of the object and the event class are assigned to basic. As a result, adjectives that do not behave as prototypical basic adjectives (that is, that are not gradable, do not act as predicates, or do not modify nouns pre-nominally) end up in the basic class in the Gold Standard. This could explain the confusability of object, BO, and basic adjectives using contextual cues.

As for the distinction between basic and event adjectives, both levels show difficulties, although of a different kind. Level *morph* correctly identifies 25 event adjectives based on the mapping between between morphology and semantics. However, it classifies 11 basic adjectives as event. These are all deverbal adjectives, for instance *conservador* (‘conservative’), *intel·ligent* (‘intelligent’), or *terrible* (‘terrible’). They represent typical mismatches arising from a strongly morphologist classification: the deriving verbs do not exist in Catalan (**terrir*) or they do exist but the adjective has acquired a meaning different to the one predictable from the suffix. For instance, *intel·ligir* has a similar meaning to ‘understand’, and *conservar* means ‘to preserve’. The suffixes are no longer active, so that the adjectives denote a plain attribute.

Two of these adjectives, *insuficient* (‘insufficient’) and *responsable* (‘responsible’), are correctly assigned to class basic by level *all*. For the remaining 9 adjectives, contextual cues are enough to place them into BE, so that their “basichood” is acknowledged; however, they do not identify them uniquely as basic. As discussed in Section 6.2.1, the distributional differences between basic and event are not robust.

Level *all* further classifies 23 event adjectives as BE. In fact, this level classifies very few adjectives as event (10, contrasting with 37 event adjectives in the Gold Standard). This is also consistent with the fact that the distributional differences between basic and event are not robust, so that most event adjectives are classified both as event (because of morphological properties) and basic (because of distributional properties).

Most of the event adjectives classified as BE by the *all* level are deverbal adjectives, and thus are correctly classified by the *morph* level. They mostly involve adjectives deriving from stative verbs, such as *abundant* ('abundant'), *imperceptible* ('imperceptible'), *preferible* ('preferable'), or *revelador* ('revealing'). We have argued in Chapter 4 that deverbal adjectives deriving from stative verbs are more similar to basic adjectives than those deriving from process-denoting verbs. These cases strongly contrast with the 7 event adjectives correctly identified by the *all* level, which are mostly derived from process-denoting verbs: exportador ('exporting'), motor ('motive | motor'), receptor ('receiving'), regulador ('regulating'), resultant ('resultant'), salvador ('rescuing'), variable ('variable').

The overall number of mistakes made by both levels is almost the same, namely, 86 (*morph*) and 89 (*all*). However, as shown in the discussion the mismatches are qualitatively quite different. The default morphological mapping works well in most cases, but is of course insensitive to shifts in meaning. Contextual cues add sensitivity to these shifts, but are not enough to achieve a better representation because of two main reasons: first, the lack of robust distributional differences between event adjectives deriving from stative verbs and basic adjectives. Second, the fact that both the basic and the object classes include nonprototypical items, which are more prone to confusion.

Adjectives that are difficult for machine learning algorithms could coincide with those that are controversial for humans. Recall from Chapter 4 that we have defined a measure of controversy, namely, the entropy of the responses. Several t-tests were performed to compare the entropy of the lemmata for which the different linguistic levels made mistakes with those correctly classified. No significant relationship was found: the lemmata that cause difficulties to humans are different from those that cause difficulties to the machine learning algorithm. No relationship was found with frequency of a given adjective either, so that adjectives from different frequency bands are equally likely to be misclassified.

6.4 Discussion

This chapter completes the set of experiments devoted to the automatic acquisition of semantic classes for adjectives performed within the PhD. The major differences with respect to the experiments explained in Chapter 5 are summarised in what follows.

We have tackled the acquisition of information regarding polysemy in viewing polysemous class assignments as an instance of multi-label classification. We have developed an architecture that corresponds to the most typical approach to multi-label problems in Machine Learning, namely, to split the task into a series of binary decisions (basic or not, event or not, object or not) and subsequently merge the individual decisions into a full classification. Thus, polysemy is implicitly acquired.

In Experiment A (Section 5.1), we treated polysemous classes as if they were distinct classes, at the same level as monosemous classes. That approach was clearly not adequate, because polysemous classes are not homogeneous (different members exhibit different profiles depending on their most frequent sense) and are not clearly differentiated from monosemous classes (polysemous adjectives share some aspects of their behaviour with the monosemous classes they participate from). In addition, polysemous classes are smaller, so that in any supervised effort, sparser evidence is available for polysemous as opposed to monosemous classes. In the present approach the aspects shared by the polysemous adjectives with each monosemous class can be separately accounted for, which was not possible in Experiment A.

In the present experiment, some types of polysemy are recognised when combining features, but the two best levels of description overgenerate polysemy and do not properly distinguish between monosemous and polysemous adjectives. When using contextual features, the most plausible explanation is as follows. In the binary decisions (basic or not, event or not, object or not), polysemous items present feature values that are closer to those of other classes, so that they lower or raise the threshold to a value that includes monosemous adjectives. Despite these problems, Experiment C captures polysemy in a more successful fashion than Experiment A, so that progress has been made in the acquisition of polysemy.

The acquisition of semantic classes has been performed separately on several levels of linguistic descriptions using several definitions. Morphological, syntactic and semantic properties have been separately encoded, and a systematic comparison in terms of accuracy has been performed. To properly perform the comparisons, the architecture includes 10x10 cross validation, so that 100 different accuracy estimates are available for each experiment for each level. We have used a corrected resampled t-test standardly used in Machine Learning to test the significance of the differences between the accuracies in the binary decisions regarding class membership.

As for the full classification, 10x10 cv yields 10 different full class proposals for each adjective. The same test has been applied to test the significance of the differences in accuracy for the full classification including polysemy information. We have provided three different accuracy measures: full, partial, and overlapping, corresponding to the three definitions of agreement established in Chapter 4. We have argued that the partial measure gives an idea of the usefulness of the classification, because it is useful to take into account the degree of overlap with the targeted classes (for which credit is given in partial and overlapping accuracy), beyond the exact match required in full accuracy.

The best accuracy results are obtained with the *all* level. They represent an average improvement of 11.4% over the baseline measured with full accuracy, and 20.4% with partial accuracy. With this level, for 90.7% of the adjectives there is some overlap between the classes assigned by the Machine Learning algorithm and the classes in the Gold Standard, against a 65.2% baseline.

As for the comparison between levels, morphology (*morph*) clearly beats levels based on contextual information (*func, uni, bi, sem*). Among the levels that use solely contextual information, *func*, accounting for the syntactic functions of the adjective, and *sem*, which takes into account other distributional correlates of semantic properties, achieve the best results. However, only *func* is statistically better than *uni* and *bi*.

The main problem for contextual features is the distinction between basic and event, because, as discussed throughout the chapter, no distributional cues seem to be robust enough to distinguish them.

In discussing the web experiment reported in Chapter 4, we have found that event adjectives are also highly problematic for humans, in the sense that most variation is observed in the classifications provided by human judges for adjectives in this class. Aspects such as the *Aktionsart* or aspectual class of the deriving verb plays a role in the “eventhood” of the adjective. Adjectives deriving from stative verbs show meanings and linguistic behaviour that are closer to the basic class. In contrast, adjectives deriving from process-denoting verbs have a more eventive interpretation.

Another source of variation within event adjectives is the suffix. Participles usually denote resulting states, adjectives derived with *ble* are have a passive kind of argument structure (their argument is the internal argument of the verb), and adjectives bearing suffixes *or* and *iu* are active in their meaning. As an example, consider *construït*, *construïble*, and *constructor*, ('built', "buildable", 'building/builder'), all derived from verb *construir* ('build').

In the adjective database used for the experiments, 8 deverbal suffixes are distinguished. In contrast, 23 denominal adjectives are included. Denominal suffixes, despite their much higher number, show much less variability with respect to the semantic class of the adjectives they build than deverbal suffixes. To this respect, note that in the web experiment 3 different patterns were defined for the event class, as opposed to a single pattern for the basic class and a single one for object class.

To sum up, both in experiments involving human annotation and in machine learning experiments, the event class strikes as problematic. The event class is not homogeneous, for the reasons outlined, and at least some members do not exhibit properties that distinguish them from the basic class. As a result, it is difficult to distinguish from the basic class both in terms of semantic and syntactic or distributional properties.

The morphology-semantics mapping is the most robust clue to the targeted classification. A very strong deviation from the etimologically expected meaning is needed for the adjective to exhibit sufficiently rich cues for contextual features to improve on the results obtained with *morph*. These cases are the ones that cause level *all* to show a slightly better average accuracy.

Note that these results are contradictory with the results in Experiment B. In Section 5.2.4.3, a test of a classification based on morphology and the clustering solution based on distributional features yielded lower scores for morphology ($p_o = 0.65$, equivalent to 65% accuracy, and $K = 0.49$) than for the clustering solution ($p_o = 0.73$, or 73% accuracy, and $K = 0.56$).

In a further test using C4.5 on the same data (Boleda, Badia and Schulte im Walde, 2005), adding the *tuning subset* to the proper Gold Standard (so that 180 lemmata were used for the experiments), the results patterned with those of Section 5.2.4.3, not with those obtained in this chapter.

Recall that in Gold Standard B, the distinctions were only made between the main classes basic, event, and object. We performed 10x10 cv directly on the 3-way distinction, with no feature selection. We tested levels *morph*, *uni*, *bi*, and *func*. The most frequent baseline resulted in 46.8% accuracy. Although all levels performed significantly better than the baseline, level *func* outperformed *morph* by an average of 3.7% (73.8% vs. 70.1%). Neither *uni* nor *bi* outperformed *morph*. Level *all*, obtained by blindly combining features from all levels (totalling 176) with no feature selection, slightly outperformed *func*, but, as is the case in the experiments explained in this chapter, no dramatic jump in improvement was obtained.

Table 6.24 (reproducing data in Boleda, Badia and Schulte im Walde (2005)) lists the mean and standard deviation of the accuracies obtained for each level.[8]

We conclude that the difference in performance is due to differences in the data, caused by differences in the methodology to build the two Gold Standard sets. Recall from Section 4.1 that Gold Standard B was randomly chosen in a token-wise fashion. For Gold Standard C,

[8] No standard deviation is given for the baseline, because it was computed on the whole data.

Level	Accuracy
bl	46.8
morph	70.1±0.3
uni	68.8±0.6
bi	67.4±0.8
func	73.8±0.3
all	74.7±0.5

Table 6.24: *Results of previous comparison between levels.*

as explained in Section 4.2.1, a stratified sampling approach was followed that accounted for variability in frequency, derivational type, and suffix. As a result, denominal and participial adjectives are underrepresented in Gold Standard C, while not derived and deverbal adjectives are overrepresented. This sampling approach was consciously followed, so as to have greater variability in the kinds of phenomena covered in the Gold Standard.

Table 6.25 shows that only 7.1% of the lemmata in Gold Standard C are participles, while they represent 22.6% of the adjectives in the database. Conversely, 26.2% of the lemmata in Gold Standard C are deverbal, and they represent 17.4% of the database.

Deriv. type	#DB	%DB	#GS	%GS
denominal	860	37.5	70	33.3
not derived	515	22.5	70	33.3
participial	517	22.6	15	7.1
deverbal	399	17.4	55	26.2
total	*2291*	*100*	*210*	*100*

Table 6.25: *Derivational type in adjective database and Gold Standard C.*

In Experiment B and Boleda, Badia and Schulte im Walde (2005), the algorithms modeled the event class in terms of the properties of participles, because they were the most frequent type of event adjectives and the remaining event adjectives do not present homogeneous properties. In Experiment C, the lack of distributional properties common to the event class is most clearly shown, because of the low frequency of participles. Therefore, it is not at all clear that level *morph* is the best level for our task: it accounts better for the event class, but then, event adjectives that are not participles are the least frequent type of adjective.

In addition, the error analysis performed for Experiment C has shown that, although the number of mistakes made with level *morph* and *all* is comparable, the kinds of mistakes are qualitatively very different. Level *all* has trouble identifying event adjectives, as could be expected because of the lack of robust distributional clues, and also discriminating between basic and object. However, it captures shifts in meaning to a larger extent than the *morph* level. Level *morph* provides a quite uniform mapping: from deverbal and participial to monosemous event (so that no shifts in meaning are accounted for), and from denominal to BO (so that it overgenerates this type of polysemy). This level does not identify a single object adjective.

Because a quite different configuration is obtained with level *all*, we conclude that it is not the optimal way to combine the strengths of each level of description. An alternative would be to build an *ensemble classifier*, a type of classifier that has received much attention in the Machine Learning community in the last decade and continues being an active area of research (Dietterich, 2002; Witten and Frank, 2005).

When building an ensemble classifier, several class proposals for each item are obtained, and one of them is chosen on the basis of majority voting, weighted voting, or more sophisticated combination methods. It has been shown that in most cases, accuracy of the ensemble classifier is higher than the best individual classifier (Freund and Schapire, 1996; Breiman, 1996; Dietterich, 2000; Breiman, 2001). The main reason for the success of ensemble classifiers is that they gloss over the biases introduced by the individual systems.

Several methodologies for the construction of ensemble classifiers have been proposed. Among the most used are methods for independently building several classifiers, by (a) varying the training data or (b) the subset of features used, (c) manipulating the labels of the training data, or (d) introducing random noise. See Dietterich (2002) for further details and references on the different methods.

In our setting, the simplest way to build an ensemble classifier would be to use the different levels of description as different subsets of features, that is, to use method (b). Naively viewed, this would be as having a team of linguists and NLP engineers, each contributing their knowledge on morphology, on n-gram distribution, on semantic properties, etc., and have them reach a consensus classification.

In related work, Rigau et al. (1997) combine several heuristics for genus term disambiguation in MRD dictionaries (Spanish and French). The improvement of a simple majority voting scheme over the best single heuristic was of 9% and 7% accuracy, for Spanish and French, respectively. They achieved an overall 80% accuracy. On a different approximation, van Halteren et al. (1998) build an ensemble classifier for part of speech tagging by combining the class predictions of different algorithms.

The two systems exhibit a *gang* effect (van Halteren et al., 2001) by implementing several voting schemes. An alternative approach is to explore the *arbiter* effect, by which a second-level machine learning algorithm is trained on the output of the several individual classifiers. van Halteren et al. (2001) test different system combinations and voting schemes for POS-tagging on three tagged corpora. For all tested datasets in this piece of research, the ensemble classifiers improve upon the best component tagger. The error reduction rate is 11.3% to 24.3% depending on the corpus.

We have performed a preliminary test that indicates that ensemble classifiers can indeed offer fruitful results for our task. The test involves establishing 10 different majority classifiers by taking the mode (most voted class) of each run, testing several combinations of levels. For the approach to work, at least three levels have to be combined. The results of the test for some combinations of 3 to 6 levels of description are depicted in Table 6.26, which also includes the accuracies obtained with the zero rule baseline (*bl*) and the best single level (*all*) for comparison.

In any of the combinations tested, accuracy improves over 10%. The best result is obtained when combining all levels of description, and jumps to a mean 84% (full accuracy) or 91.8% (partial accuracy). Note that with this procedure 95.7% of the classifications obtained with the ensemble classifier present some overlap with the class assignments in the Gold Standard (see overlapping accuracy). These results represent a raw improvement in terms of full accuracy of 33% over the baseline and 21.7% over the best single classifier.

The accuracies obtained correspond to the simplest possible ensemble classifier, obtained by unweighted voting over the classifications obtained with each feature set. This simple test, however, provides a strong indication that ensemble classifiers are a more powerful and ade-

Levels	Classifs.	Full Acc.	Partial Acc.	Overl. Acc.
morph+func+uni+bi+sem+all	6	84.0[9]	91.8[10]	95.7[11]
morph+func+uni+bi+sem	5	82.3[12]	91.3[13]	95.9[14]
func+uni+bi+sem	4	81.5[15]	91.0[16]	95.9[17]
morph+func+sem+all	4	72.4[18]	83.6[19]	89.3[20]
morph+func+sem	3	76.2[21]	85.8[22]	90.6[23]
bl	-	51.0±0.0	60.5±0.0	65.2±0.0
all	-	62.3±2.3	80.9±1.5	90.7±1.6

Table 6.26: *Results for ensemble classifier.*

quate way to combine the linguistic levels of description than simply merging all features for tree construction.

A preliminary version of the experiments presented in this chapter has been published in the following article:

Boleda, G., Badia, T., and Schulte im Walde, S. (2005). Morphology vs. syntax in adjective class acquisition. In *Proceedings of the ACL-SIGLEX 2005 Workshop on Deep Lexical Acquisition*, pages 1119–1125, Ann Arbor, USA.

Chapter 7
Conclusion

7.1 Conclusions

In this thesis we have pursued a line of research that seeks to induce semantic classes for adjectives from morphological, syntactic, and semantic evidence.

The definition of a broad semantic classification for adjectives is a controversial matter from a theoretical point of view. A major effort in the thesis has been to combine data from literature review with several empirical approaches to adjective classification, so as to test the strengths, weaknesses and possibilities of the classification proposals.

Three main sources of empirical insight have been exploited. First, the differences observed in the manual annotation of adjectives by several judges (agreement analysis). Second, the distribution of feature values across adjective classes. Third, the results of unsupervised and supervised machine learning experiments applied on the manually annotated data using the defined features.

Throughout the thesis, we have devoted every effort to analyse and interpret the pieces of information gathered, and we have revised our hypotheses and re-interpreted the data according to the results of these analyses. Thus, the techniques and tools we have used have served as tools for linguistic research.

In the Introduction, we have stated three goals for this thesis. We now review each of the goals and summarise the findings obtained with respect to each of them.

Goal 1: To define a broad, consistent, and balanced semantic classification for adjectives
The definition of a semantic classification for adjectives is a controversial matter from a theoretical point of view.

We have proposed an initial classification into qualitative, relational, and intensional adjectives, based on proposals in formal semantics and descriptive grammar. The analysis of the human annotation data and the results of an unsupervised experiment has led us to revise the classification by dropping the intensional class and adding a new class. The revised classification distinguishes between basic (formerly, qualitative), object-related (formerly, relational), and event-related adjectives. It is backed up by the treatment of adjectival semantics in Ontological Semantics.

The classes can be characterised as follows.

- Basic adjectives typically denote attributes. In Catalan, they can be used nonrestrictively (in pre-nominal position), predicatively, and tend to be gradable. They frequently enter into coordinating constructions and can appear further away from their head noun than other types of adjectives. In general, they are characterised by exhibiting the greatest variability in terms of syntactic function and distributional properties. The core members of this class are lexically organised in terms of scales, and have antonyms. However, most basic adjectives do not have antonyms in the narrow sense of the term.

- Event adjectives denote a relationship to an event. Because most event adjectives are deverbal, they tend to have a more complex argument structure than basic adjectives, and in particular, to bear complements. Probably for this reason, they tend not to prenominally modify nouns. Some of them enter into predicative constructions with a higher frequency than basic adjectives.

- Object adjectives denote a relationship to an object. They are frequently used to point to a subtype of the class of objects denoted by their head noun. In Catalan, object adjectives appear rigidly after the noun. They cannot be used nonrestrictively, and can act as predicates under very restricted circumstances. Most of them are not gradable and do not bear complements. The typical noun modified by an object adjective is feminine and appears in a definite NP, which in turn is embedded within a PP. Although other kinds of adjectives appear in these environments, object adjectives do so in a much higher proportion of cases.

Most of the properties identified in the feature analyses carried out in this thesis coincide with the parameters explored in the literature. However, our approach has uncovered several pieces of information that are not usually discussed in linguistic theory, because they are easier to uncover from corpus statistics than from introspection, the main methodology currently used in linguistic analysis. The main contributions concern the behaviour of the object class with respect to definiteness and the syntactic function of the head noun. Further linguistic research is needed to provide an explanation for the pieces of data described above.

We have further shown that although polysemous adjectives share different aspects of the core classes to which they belong, they do not form a homogeneous, distinct class. Two main aspects play a role in the difficulty to characterise polysemous adjectives. First, the fact that most adjectives have a more frequent sense in the corpus, so that their behaviour corresponds to one or the other classes involved. Second, the fact that while there are some clear cases of polysemy, for some other cases rather ambiguity seems to be at play.

The classification proposed presents some futher difficulties, which mainly concern the characterisation of the event class. The event class arose as a result of the unsupervised experiments presented in Section 5.1, because the semantics of some of its members, particularly those built with suffixes *ble*, *or*, and *iu*, does not easily fit into the basic class. It is also backed up by proposals within Ontological Semantics, and to some extent in WordNet.

However, the results reported in this thesis indicate that it cannot be viewed as a compact class. Probably, the event class consists of distinct subclasses, that correspond to a combination of suffix type and aspectual class of the deriving verb. Also, some of its members (most notably, those derived from stative verbs) are better placed in the basic class, because they can be argued to denote plain attributes, instead of being related to an actual event.

Object adjectives are the most compact class, but they also exhibit variability. Many denominal adjectives, such as nationality or ideology adjectives, are ambiguous between the basic and the object class. Some others, such as *fangós* (‘muddy’) or *diari* (‘daily’), have an object component but a much more specified relationship to the object than is usually the case for object adjectives.

As for basic adjectives, the core of the class does exhibit a set of common semantic and distributional properties, but many adjectives that are assigned to this class in the several Gold Standards do not share these properties. Among the types treated in the literature, intensional adjectives, modal adjectives, so-called adverbial adjectives have idiosyncratic properties. Other kinds of adjectives that are not usually tackled in the literature do not straightforwardly denote attributes and do not share many of the properties of basic adjectives.

This discussion raises the issue of whether it makes sense to attempt at a global classification of adjectives in the fashion pursued in this thesis. A possible answer to this question is that a broad classification serves limited purposes, and that necessarily many subtypes of adjectives do not fit in the definition of the main classes. Probably, the best way to integrate these subtypes into the classification would be to build a hierarchy, instead of a flat classification like the one we have proposed here. We now turn to the findings regarding the second goal of the thesis.

Goal 2: To test the feasiblity of the classification task by humans Within the thesis, three different manual annotation experiments have been carried out. The main contribution has been the design of a web experiment to gather large amounts of data for each adjective. Web experiments are a promising way to gather linguistic data, which has not been much exploited within the linguistic or NLP communities. In Psychology and Psycholinguistics, in contrast, they are increasingly used.

In the web experiment, we have gathered data from 322 judges, each classifying 30 out of the 210 adjectives in the Gold Standard. Traditionally, agreement studies have been performed for a small number of judges and a relatively large number of items to be classified. Our situation is the reverse, which has motivated the development of a specific approach to measuring agreement.

We have used previously defined descriptive measures (proportion of observed agreement and proportion of specific agreement) as well as a measure that corrects for chance agreement, kappa. We have argued that confidence intervals should be estimated for any measure of agreement used, and we have proposed a sampling approach to obtain robust estimates in experimental tasks with a large number of judges.

The web experiment was aimed at the acquisition of semantic classes including polysemy information, so that we allowed the judges to provide multiple class assignments for a single adjective. Weighted measures of agreement, such as weighted kappa, are called for to adequately model partial agreement in multiple class assignments.

We have proposed a weighting scheme based on an explicit model of the task to be performed by the humans. Even if this model is oversimplified, by comparing the weighted scores with the lower and upper bounds for the agreement values we have shown that it sensibly accounts for partial agreement. The best estimate for agreement among human judges as assessed on the web data is a K value of 0.31 to 0.45. This value is too poor for academic standards, so that we have established the final classification for the dataset relying on a committee of 3 experts.

A comparison of the data obtained through the web experiment with the expert classification has shown systematic deviations from the intended classification, which indicates problems in the design of the task. Through the analysis of agreement results we have also uncovered difficulties in the classification, particularly regarding the event class, as has already been mentioned above. This analysis has been possible by using entropy as a measure of intra-item agreement.

To sum up, we have used the process of building a Gold Standard as a further empirical approach to the overall goal of establishing a meaningful and adequate semantic classification for adjectives. However, we have failed to produce a reliable Gold Standard including polysemy information. Next section suggests ways to improve upon the methodology developed thus far.

Goal 3: To test the feasiblity of the classification task by computers We have performed three machine learning experiments to automatically classify adjectives into semantic classes. The results obtained indicate that the semantic classification of adjectives using morphological and distributional information is feasible to a large extent, at least for Catalan. The classification proposal and the methodology could be straightforwardly extended to other Romance languages, such as French, Spanish, or Portuguese.

The first two experiments have been directed at testing the initial classification proposal, revising it, and providing an initial characterisation of each class. These purposes have motivated the use of an unsupervised technique used, clustering. In the second clustering experiment, performed to distinguish between the main class of adjectives (ignoring polysemy), the estimated accuracy obtained is 73%, compared to a 49% baseline.

The acquisition of polysemy has been tackled with a supervised technique, Decision Trees. Viewing the problem as a multi-label classification task, we perform the classification in two steps: first, binary decision on each of the classes (basic or not, event or not, object or not). Second, merging of the decisions to achieve a full classification.

A specific goal of the experiment was to test the strengths and weaknesses of several linguistic levels of description (morphology, syntax, semantics) for the task at hand. We have shown that in general terms, features defined on the basis of contextual information correctly account for object adjectives, but they cannot distinguish between basic and event adjectives. Morphology offers a default classification that fares relatively well with all classes, but cannot account for shifts in meaning that cause polysemy or change of class. Although contextual features are more sensitive to these deviations from the expected class, they often do not distinguish between polysemy and class shift.

The best results are obtained with a combination of features from different levels of linguistic description, which obtains accuracy estimates of 62.3%, under the strictest evaluation conditions. These results represent an average improvement over the baseline of 11.4%.

The error analysis has suggested that using all levels in a single experiment is not the best way to combine the strengths of each level of description. Ensemble classifiers, which build a classification out of several proposed classifications, are an attractive alternative. Majority voting over the classifications yielded by each individual level raises accuracy to a mean 84% (full accuracy), which represents a raw improvement of 21.7% over the best single classifier, opening the way to richer ways to combine different types of linguistic evidence.

7.2 Future work

We suggest three main lines for future research. The first one is to explore several extensions within the same approach to the task and to provide external evaluation. The second one is to consider alternative information for the acquisition of semantic classes, most notably selectional restrictions. The third one is to investigate the theoretical implications of the results for linguistic research. We now outline each of these three lines of research.

We begin with extensions in the type of linguistic information taken into account. Within the morphological and syntactic levels of description, some potentially useful pieces of information have not been exploited in this thesis. As for morphology, in future work, it would be advisable to exploit prefixation in addition to suffixation. For instance, we have informally observed that participles with an *in-* prefix (equivalent to English *un-*) tend to be more basic-like than those without the prefix. In fact, Bresnan (1982) used prefixation with *-un* as a criterion to identify adjectival uses of participles. Also, property-denoting nouns are built out of adjectives with suffixes such as *-itat*, *-esa* ('-ity, -ess'). It can be expected that mostly basic adjectives are subject to these morphological processes.

As for syntax, coordination has not been exploited, beyond its use as feature in the n-gram models. In Chapter 3 we have shown that basic and object adjectives do not coordinate. It should be explored whether event coordinates with basic or object adjectives. As explained in Section 3.7, Bohnet et al. (2002) apply a bootstrapping approach using order and coordination information to classify German verbs. A similar approach could be devised for Catalan.

At the interface between syntax and semantics would be the acquisition of argument structure for use in semantic classification. The argument type of an adjective (e.g., proposition or entitiy) tends to correlate with a particular subcategorisation type (e.g., clause or NP). The strategy developed in Yallop et al. (2005) for the acquisition of syntactic subcategorisation patterns for English adjectives could be adapted to Catalan and further extended to the acquisition of argument structure. However, Yallop et al. (2005) use a robust statistical parser, a resource that is currently not available for Catalan. Alternative methods using shallow parsers, which are available for Catalan (CatCG, the tool used here; FreeLing, an open-source tool that provides chunking and dependency parsing; see Atserias et al. (2006)), should be explored.

The results of the experiments have been based on a 14.5 million corpus, containing formal written text. This corpus is quite small compared to corpora standardly used in other Lexical Acquisition tasks. For instance, the *British National Corpus* (Burnage and Dunlop, 1992), used in acquisition tasks for English such as Lapata (2001), contains 100 million words. The results should be checked against larger and qualitatively different corpora. For Catalan, a large web corpus has been recently developed within GLiCom (Boleda et al., 2006), which could be used for the acquisition experiments.

A third type of extension is the use of other machine learning techniques. The conclusions of this study have been reached on the basis of two techniques, namely, clustering and Decision Trees. More specifically, a detailed analysis has only been provided for the k-means and C4.5 algorithms. The feature exploration and the error analyses performed make it plausible that the conclusions reached would extend to other machine learning techniques. However, machine learning approaches based on a different modeling of the learning problem should be tested to confirm this hypothesis, such as k Nearest Neighbours, Bayesian classifiers, or Support Vector Machines, a technique that has recently proven successful for many NLP tasks.

However, we have shown that for our task, a combination of classifiers built on the basis of several levels of linguistic description outperforms the best single set of features; by extension, it should outperform the best single machine learning algorithm. We have barely scratched the surface of the possibilities of ensemble classifiers, as we have used the simplest approach, unweighted majority voting. Stacked classifiers and different types of voting schemes, as explored in, e.g., van Halteren et al. (2001), should be tested to reach stable conclusions with respect to the use of ensemble classifiers for our task.

Up to now, we have only evaluated the results internally, through comparison with the Gold Standards. Evaluation setups that embed the information acquired into more complex systems should be developed to assess its usefulness. In the Introduction, we have argued that the classification obtained could be used in low-level tasks, such as POS-tagging. A statistical parser trained with and without class information is a suitable test for this approach.

Evaluation could also be performed by using information on semantic class for more sophisticated tasks involving extraction of semantic relationships. For instance, object adjectives can evoke arguments when combined with predicative nouns (*presidential visit* - a president visits X). For projects such as FrameNet (Baker, Fillmore and Lowe, 1998), these kinds of relationships can be automatically extracted for adjectives classified as object. For event adjectives, the reverse applies, as the adjective can be deemed equivalent to a verbal predicate: *flipping coin* - a coin flips. This kind of information can be applied to tasks that have recently received much attention within the NLP community, such as Paraphrase Detection and Textual Entailment Recognition. The knowledge gathered through these tasks can be applied to, e.g., summarisation: the noun-adjective constructions are shorter than their sentential counterparts.

In this thesis, we have used morphological, syntactic, and semantic information for classification. A major piece of information that we have not yet explored are selectional restrictions. Including this type of information is a second line of research arising from this thesis.

Selectional restrictions specify the kinds of arguments a given predicate prefers, or even allows. Some pieces of evidence in feature analysis, such as object adjectives modifying feminine heads in a higher proportion than other classes of adjectives, suggest that semantically different adjectives select for semantically different types of arguments. If successful, this level of information would make the semantic classification of adjectives largely language-independent (provided sufficiently similar semantic classes are present in the targeted language), because it would not rely on surface properties.

However, this level of information faces two major difficulties. The first one is that in the literature, no clear predictions with respect to which kinds of heads should be expected for different semantic classes are found. We have suggested that object adjectives tend to modify abstract nouns in a higher proportion of cases than other kinds of adjectives, although this hypothesis is still to be tested.

The second difficulty is the establishment of a set of the categories for the selectional restrictions. The difficulty is increased by the fact that polysemy and ambiguity, which we want to account for, affect the head nouns of adjectives just as they affect adjectives.

In research on acquisition of selectional restrictions (Resnik (1993), Abe and Li (1996), McCarthy (2001), McCarthy and Carroll (2003)), lexical resources such as thesauri or WordNet are typically used to define the categories, allowing for different levels of granularity by cutting at different points in the hierarchies. In other approaches (Reinberger and Daelemans, 2003), the categories are unsupervisedly built via clustering.

Following this latter kind of approach, a possibility would be to cluster nouns with verbal data (that is, to group nouns according to which verbs they cooccur with), and use the resulting clusters as noun classes to acquire selectional restrictions for adjectives. The classes would be used as features for classification into semantic classes. This way, an independently motivated classification would be built (based on verbs), and relationships between verbal and adjectival arguments could be explored. However, deciding on the optimal number of clusters (classes) is a difficult issue, and the interpretation of the resulting clusters in terms of semantic classes is yet more difficult.

The third line of research involves investigating the theoretical implications of the results presented here for linguistic research. First, in the definition of a classification for adjectives. The difficulties faced have been summarised in the previous section; how best to overcome them is a matter for further research. Second, in the characterisation of each class: characteristics uncovered through feature analysis should be explained from a linguistic point of view.

Third, in the definition of polysemy within our task. Aiming at acquiring polysemy implies that different senses can be distinguished and encoded. Difficulties in distinguishing ambiguity from polysemy have been highlighted in the thesis, and have arguably affected the agreement level in the manual annotation experiments. Polysemy judgments have proved very difficult to obtain in a reliable fashion.

Relatedly, a further aspect left for future research is the design of adequate experiments to build reliable Gold Standard sets for our task. In addition to problems involving the definition and detection of polysemy, we have argued that two main factors explain the poor agreement obtained in the web experiment: first, the use of naive subjects; second, problems in the definition of the task, as it involved meta-linguistic judgments.

The first problem is unavoidable if a large number of judgments wants to be obtained: even if the target are subjects with some training in linguistics, they will not all be experts in the specific problem tackled in a particular experiment. It has proven very difficult to design a classification task for naive subjects. Methodologies and designs used in psycholinguistics should be further checked to improve the design, so that more stable results are obtained. Such a design should not involve meta-linguistic knowledge, but linguistic intuitions.

A possibility would be to define linguistic contexts that correspond to senses in each of the semantic classes, and ask for plausibility ratings. However, these contexts are most probably lemma-dependent, so that the intended classification has to be defined beforehand. This method, then, could only be used to validate a previous classification built, e.g., by an expert committee, not to build a classification from scratch. Developing reliable methodologies to obtain complex linguistic information from naive subjects would be very useful for both NLP and linguistic theory.

Appendix A
Agreement results

This appendix provides detailed agreement results. In Chapter 4, we have reported results averaged over all test sets. Section A.1 reports the agreement values obtained per test set. Section A.2 graphically shows the distribution of agreement values for all pairs of each test set, for the three agreement definitions (full, partial, overlapping). Section A.3 shows the distribution of agreement values for each class.

A.1 Agreement results per test set

In Tables A.1 and A.2, for each estimate of the agreement value, the first row contains the mean ± standard deviation values for each test set. The second row shows the confidence interval for each test set.

Agr. Def.	Est.	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
full	p_o	.42±.15	.42±.12	.48±.16	.42±.15	.46±.14	.41±.19	.49±.14
		.35-.49	.36-.47	.4-.55	.35-.5	.41-.51	.33-.49	.43-.55
	K	.25±.16	.25±.11	.33±.18	.27±.15	.26±.15	.2±.18	.3±.13
		.18-.33	.2-.3	.24-.41	.2-.34	.21-.32	.12-.28	.25-.36
partial	wp_o	.64±.11	.65±.06	.68±.1	.63±.09	.65±.08	.66±.1	.68±.08
		.59-.69	.62-.68	.63-.72	.59-.68	.62-.68	.62-.7	.65-.71
	wK	.36±.19	.38±.1	.46±.16	.38±.15	.35±.14	.33±.18	.41±.13
		.28-.45	.33-.42	.39-.53	.31-.45	.3-.4	.25-.41	.35-.46
overlapping	op_o	.78±.13	.78±.11	.79±.11	.75±.1	.76±.11	.8±.1	.78±.12
		.72-.84	.73-.83	.75-.84	.7-.8	.71-.8	.75-.84	.73-.83
	oK	.48±.27	.52±.15	.6±.17	.49±.18	.46±.18	.47±.22	.53±.22
		.36-.6	.45-.59	.52-.67	.4-.57	.4-.53	.37-.56	.44-.63

Table A.1: *Overall agreement values per test set.*

Cl.	**Est.**	**Set 1**	**Set 2**	**Set 3**	**Set 4**	**Set 5**	**Set 6**	**Set 7**
B	p_s	.4±.22	.45±.21	.48±.2	.43±.24	.55±.16	.54±.2	.5±.24
		.3-.51	.36-.55	.39-.57	.31-.54	.49-.62	.45-.62	.39-.6
	K	.24±.25	.3±.22	.32±.25	.32±.25	.34±.19	.23±.21	.29±.22
		.13-.36	.2-.4	.2-.43	.2-.44	.27-.41	.15-.32	.19-.38
BE	p_s	.13±.24	.15±.25	.15±.24	.08±.18	.09±.18	.13±.22	.15±.26
		0:.27	.03-.27	.03-.28	-.02:.17	.01-.17	.02-.24	.03-.27
	K	.08±.24	.12±.25	.11±.23	.04±.14	.06±.17	.1±.2	.11±.24
		-.05:.22	0:.24	-.01:.23	-.04:.12	-.01:.14	0:.2	0:.22
BO	p_s	.18±.19	.11±.16	.1±.19	.19±.25	.14±.22	.12±.2	.21±.26
		.08-.27	.03-.19	.01-.19	.06-.32	.05-.23	.03-.21	.09-.32
	K	.07±.16	.04±.14	.05±.17	.15±.24	.05±.2	.02±.16	.15±.25
		-.01:.15	-.03:.11	-.04:.13	.03-.28	-.03:.13	-.05:.09	.04-.25
E	p_s	.34±.28	.2±.23	.32±.3	.37±.23	.23±.24	.38±.35	.21±.29
		.22-.47	.1-.3	.19-.46	.26-.48	.14-.32	.23-.53	.08-.33
	K	.28±.26	.13±.22	.23±.3	.25±.21	.17±.23	.31±.35	.15±.29
		.16-.4	.03-.23	.09-.36	.15-.35	.08-.26	.16-.47	.03-.28
EO	p_s	.06±.14	.1±.19	.21±.3	.15±.19	.09±.23	.19±.25	.08±.29
		0:.13	.01-.19	.05-.37	.05-.25	-.01:.19	.07-.31	-.1:.27
	K	.01±.14	.07±.18	.17±.3	.09±.15	.07±.23	.16±.24	.07±.29
		-.06:.08	-.01:.16	0:.33	.01-.17	-.03:.17	.05-.28	-.12:.25
O	p_s	.58±.19	.59±.11	.67±.2	.55±.18	.5±.18	.13±.24	.63±.19
		.49-.66	.54-.64	.58-.76	.46-.64	.43-.57	.03-.24	.55-.71
	K	.41±.19	.39±.14	.53±.27	.36±.21	.31±.18	.04±.25	.48±.19
		.32-.49	.33-.45	.41-.65	.25-.46	.24-.38	-.07:.15	.4-.56

Table A.2: *Class-specific agreement values per test set.*

A.2 Overall distribution of agreement scores

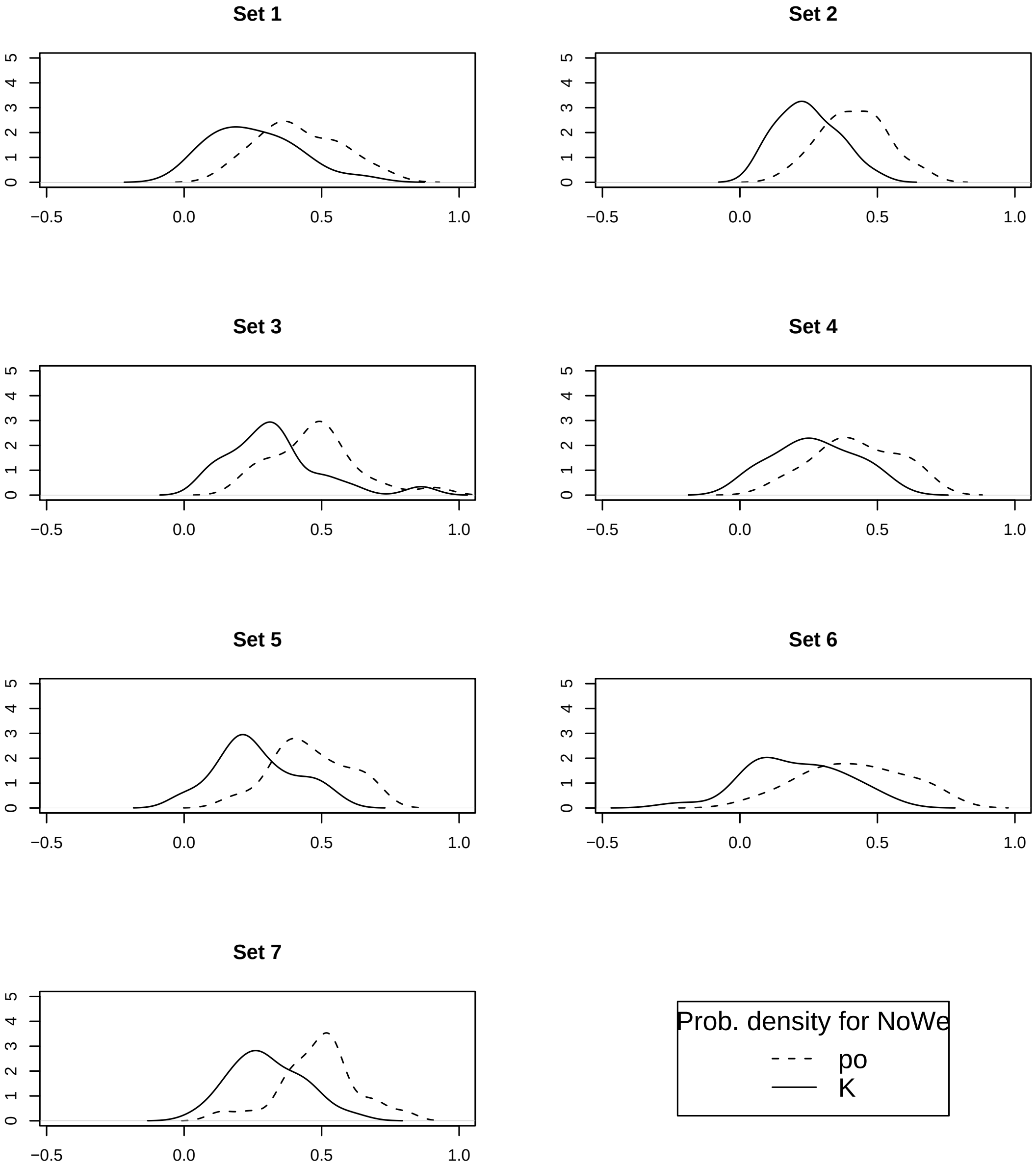


Figure A.1: *Overall distribution of agreement scores (full agreement).*

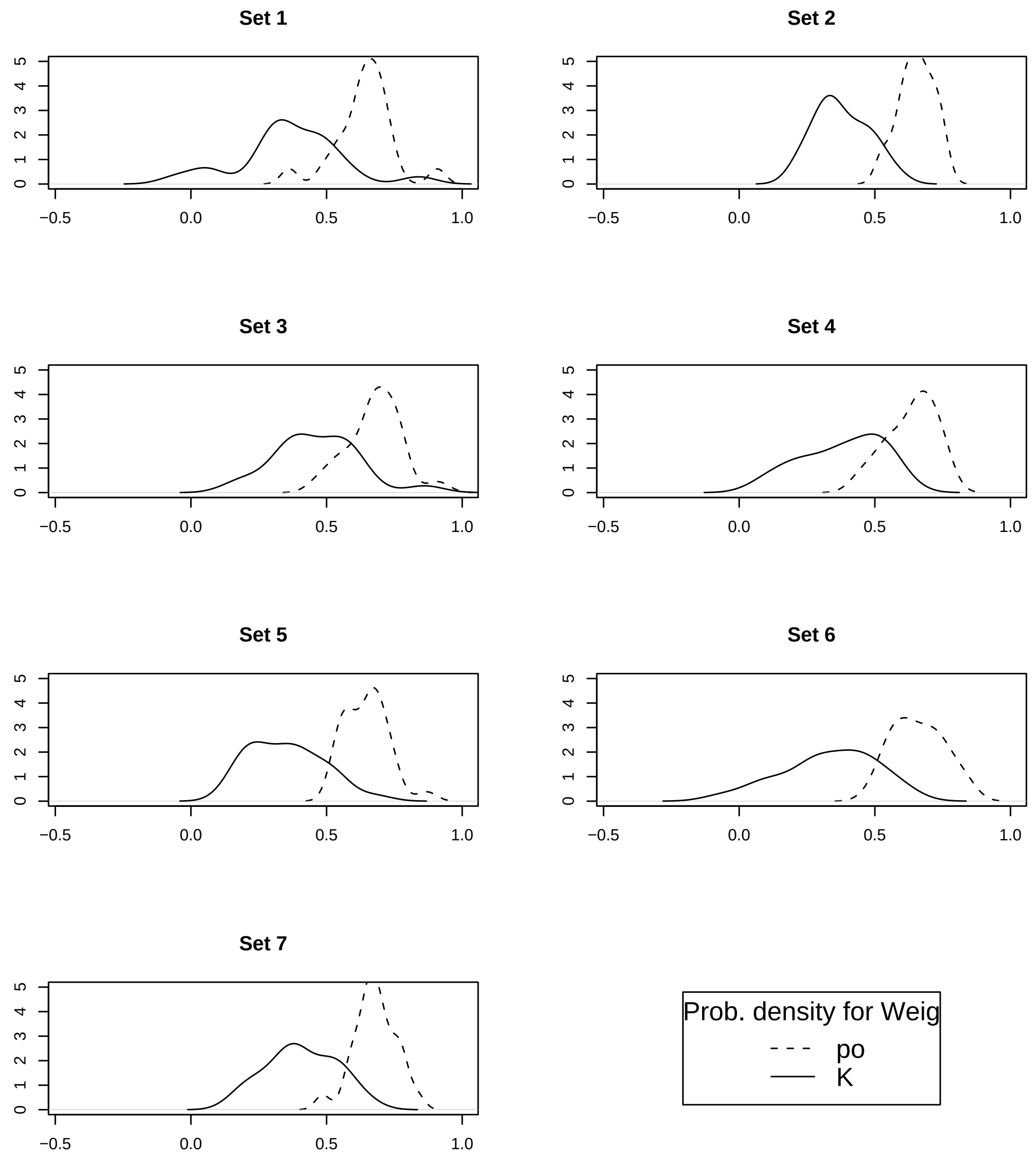


Figure A.2: *Overall distribution of agreement scores (partial agreement).*

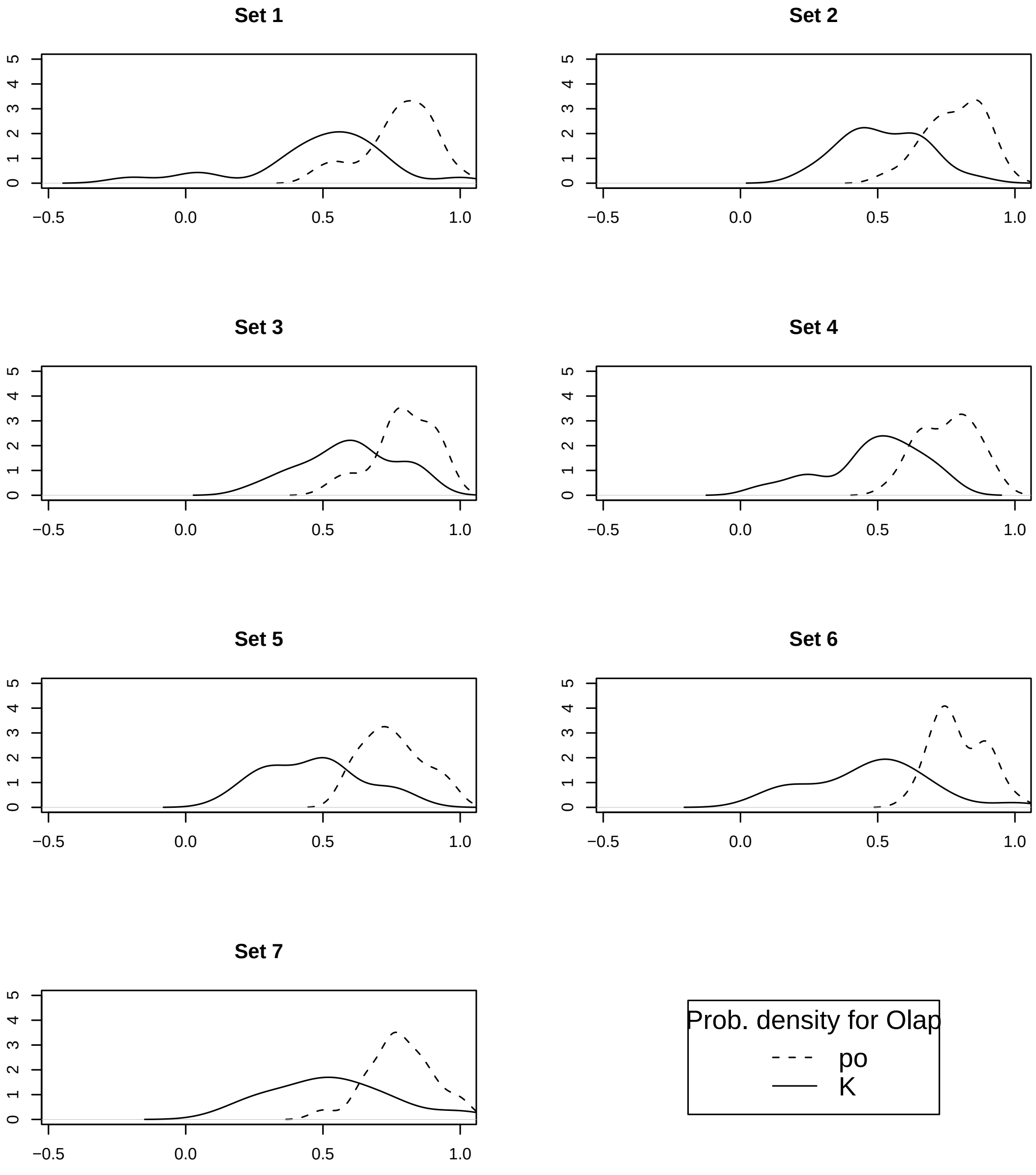


Figure A.3: *Overall distribution of agreement scores (overlapping agreement).*

A.3 Distribution of agreement scores per class

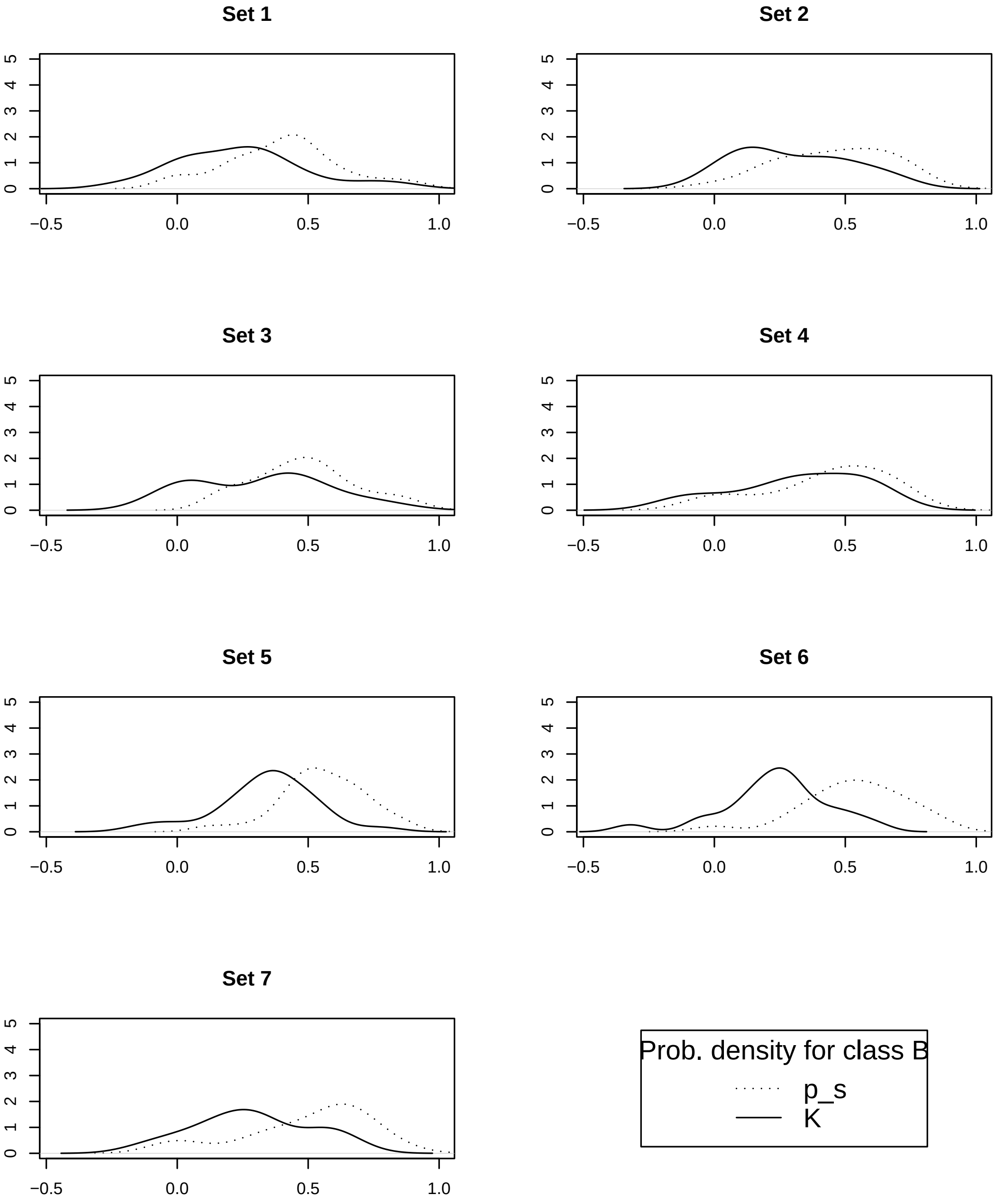


Figure A.4: *Distribution of agreement scores in class basic.*

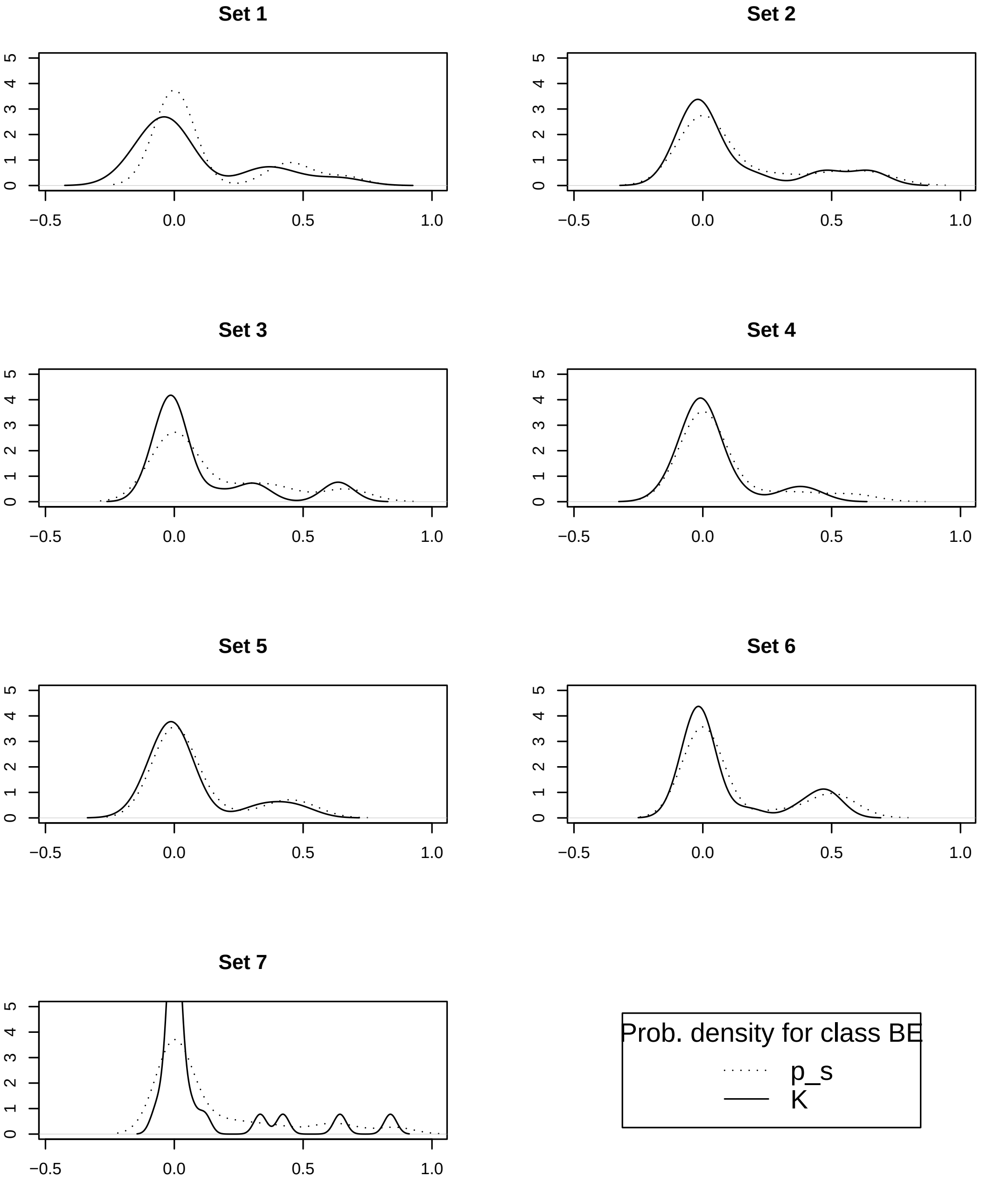


Figure A.5: *Distribution of agreement scores in class basic-event.*

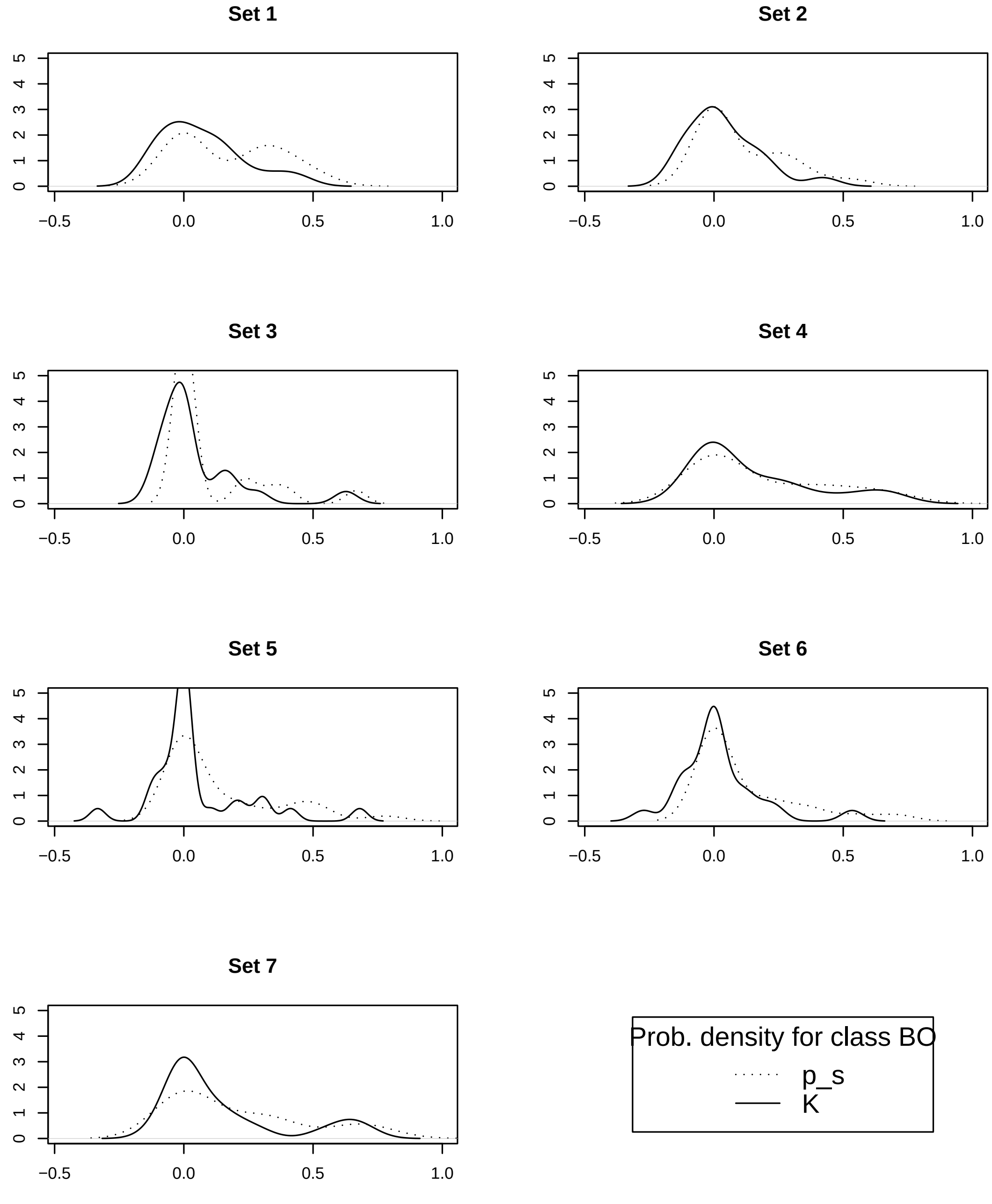


Figure A.6: *Distribution of agreement scores in class basic-object.*

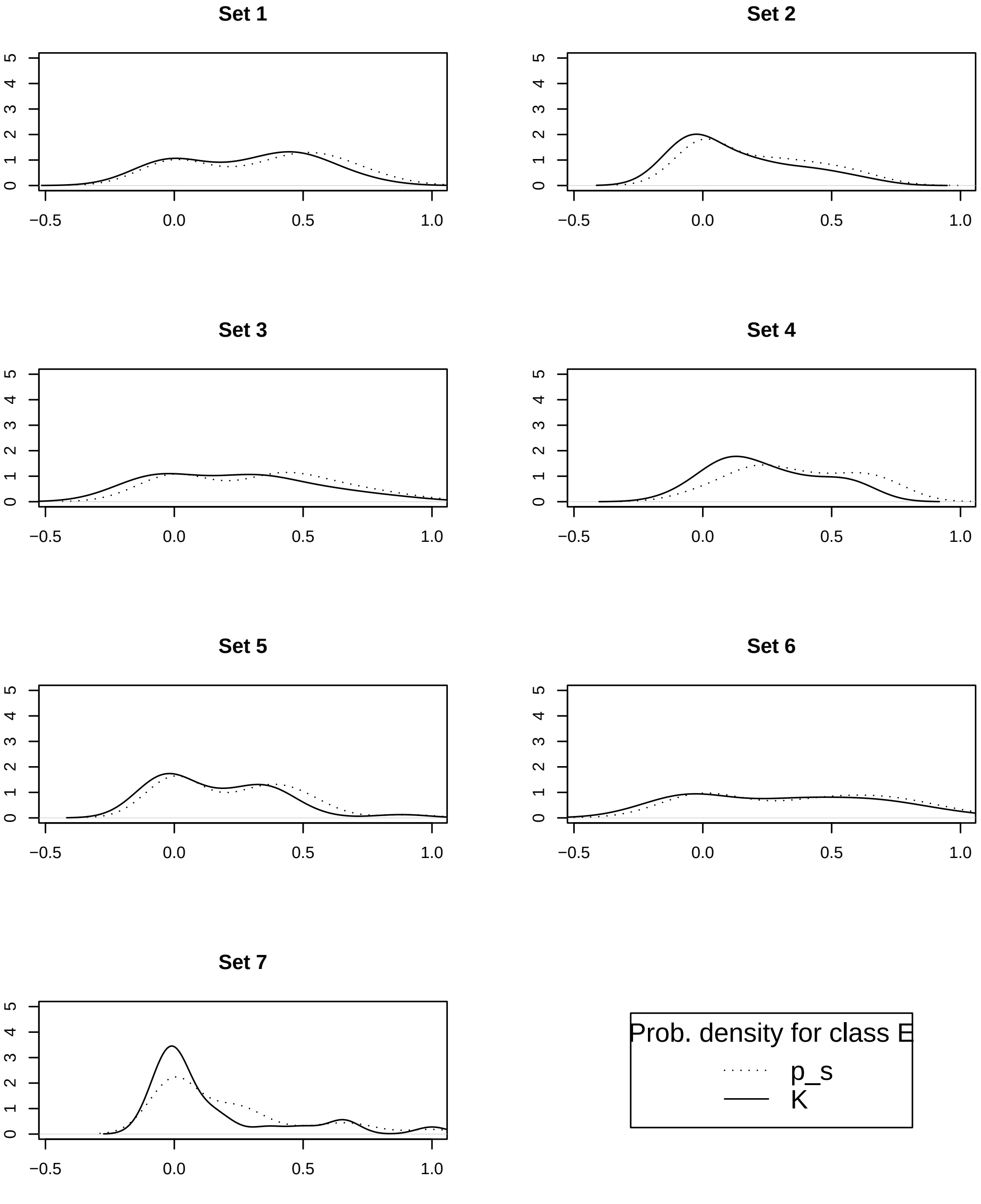


Figure A.7: *Distribution of agreement scores in class event.*

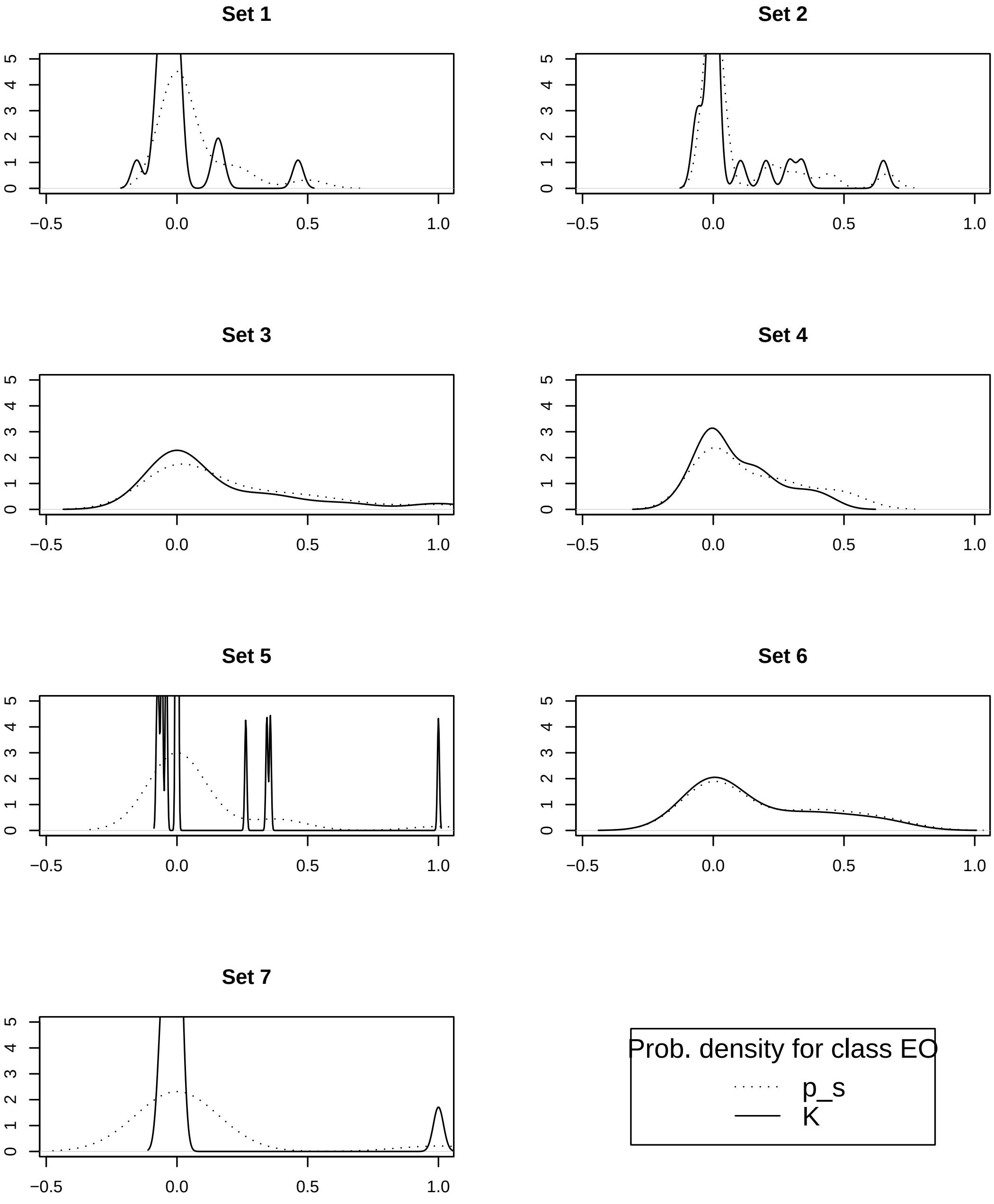


Figure A.8: *Distribution of agreement scores in class event-object.*

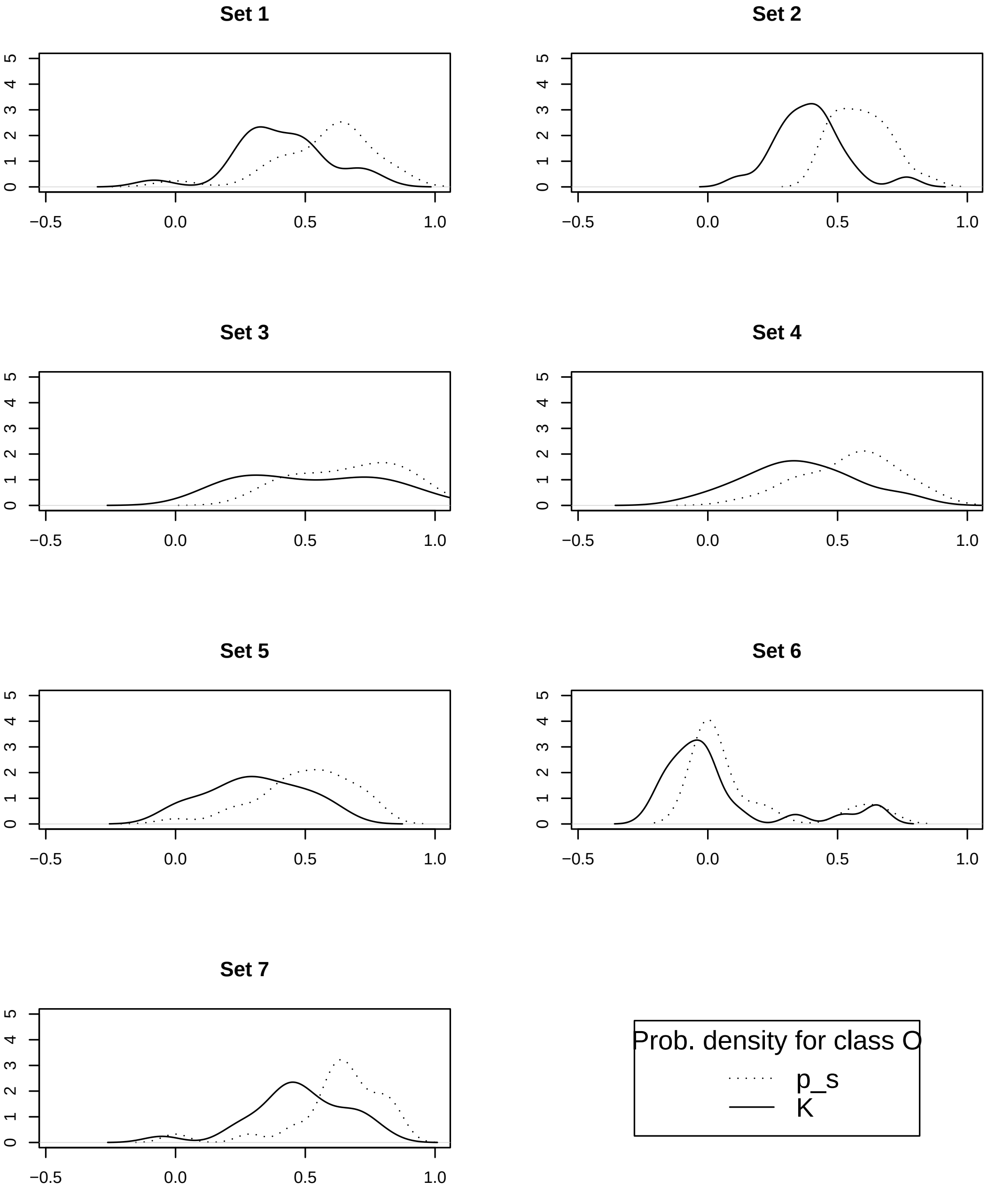


Figure A.9: *Distribution of agreement scores in class object.*

Appendix B
Data for Experiment C

Tables B.1 and B.2 contain the mean and standard deviation values of the features encoding syntactic function and semantic properties, as used in Experiment C (Chapter 6). Values for each class as well as global values for each feature are given. For each feature, one row show the mean (M) values and the following one the standard deviation (SD) values, in smaller font and italics.

Recall that semantic features *not restrictive*, *predicate (copular)*, and *predicate (other)* and are identical to syntactic features *pre-nominal mod., pred (cop.),* and *pred (other)*. Their mean and standard deviation values are only shown for syntactic features (Table B.1).

Feature	**M/SD**	**B**	**BE**	**BO**	**E**	**EO**	**O**	**Global**
post-nominal mod.	M	0.66	0.48	0.9	0.67	0.97	0.95	0.73
	SD	*0.19*	*0.19*	*0.09*	*0.22*	*0.02*	*0.05*	*0.21*
pre-nominal mod.	M	0.11	0.02	0.01	0.04	0	0.01	0.07
	SD	*0.14*	*0.02*	*0.03*	*0.08*	*0*	*0.01*	*0.12*
pred (cop.)	M	0.13	0.21	0.04	0.15	0.01	0.01	0.11
	SD	*0.11*	*0.15*	*0.03*	*0.14*	*0.01*	*0.03*	*0.12*
pred (other)	M	0.1	0.29	0.05	0.13	0.02	0.03	0.09
	SD	*0.08*	*0.14*	*0.04*	*0.13*	*0.01*	*0.03*	*0.1*

Table B.1: *Mean and standard deviation values for syntactic function features.*

Feature	**M/SD**	**B**	**BE**	**BO**	**E**	**EO**	**O**	**Global**
pred (with *estar*)	M	0.01	0.09	0.01	0.03	0	0	0.02
	SD	*0.03*	*0.1*	*0.03*	*0.06*	*0*	*0*	*0.04*
pred (global)	M	0.23	0.52	0.09	0.29	0.03	0.04	0.2
	SD	*0.16*	*0.18*	*0.07*	*0.21*	*0.02*	*0.04*	*0.18*
gradable	M	0.04	0.08	0.01	0.06	0.01	0	0.04
	SD	*0.05*	*0.09*	*0.01*	*0.08*	*0.01*	*0*	*0.06*
comparable	M	0.06	0.06	0.02	0.06	0.01	0.01	0.05
	SD	*0.06*	*0.05*	*0.02*	*0.09*	*0.01*	*0.03*	*0.07*
gradable (global)	M	0.11	0.14	0.03	0.12	0.01	0.01	0.08
	SD	*0.1*	*0.09*	*0.03*	*0.15*	*0.02*	*0.04*	*0.11*
definite	M	0.32	0.16	0.51	0.34	0.6	0.59	0.38
	SD	*0.15*	*0.06*	*0.12*	*0.21*	*0.06*	*0.15*	*0.2*
indefinite	M	0.38	0.4	0.27	0.34	0.23	0.24	0.34
	SD	*0.13*	*0.11*	*0.09*	*0.14*	*0.09*	*0.11*	*0.14*
bare	M	0.2	0.19	0.16	0.17	0.13	0.12	0.18
	SD	*0.1*	*0.1*	*0.05*	*0.1*	*0.05*	*0.06*	*0.1*
head as subject	M	0.08	0.06	0.07	0.09	0.09	0.09	0.08
	SD	*0.03*	*0.02*	*0.02*	*0.04*	*0.02*	*0.08*	*0.04*
head as object	M	0.12	0.1	0.12	0.11	0.14	0.11	0.12
	SD	*0.05*	*0.06*	*0.04*	*0.04*	*0.04*	*0.05*	*0.05*
head as prep. comp.	M	0.37	0.25	0.51	0.34	0.54	0.55	0.4
	SD	*0.13*	*0.11*	*0.09*	*0.12*	*0.04*	*0.11*	*0.14*
distance to head noun	M	0.44	0.45	0.23	0.49	0.19	0.18	0.38
	SD	*0.28*	*0.16*	*0.15*	*0.31*	*0.1*	*0.14*	*0.27*
binary	M	0.22	0.29	0.17	0.33	0.18	0.19	0.23
	SD	*0.17*	*0.24*	*0.04*	*0.2*	*0.06*	*0.07*	*0.16*
plural	M	0.28	0.28	0.27	0.32	0.4	0.31	0.3
	SD	*0.11*	*0.14*	*0.08*	*0.13*	*0.17*	*0.14*	*0.12*
feminine	M	0.49	0.39	0.5	0.5	0.39	0.56	0.5
	SD	*0.12*	*0.13*	*0.08*	*0.1*	*0.12*	*0.14*	*0.12*

Table B.2: *Mean and standard deviation values for semantic features.*

Appendix C
Gold Standard data

This appendix lists the three Gold Standard sets used for the experiments in Chapters 5 and 6. They were obtained with the methodologies explained in Chapter 4.

C.1 Gold standard A

Gold standard A consists of the 101 lemmata listed below. The list includes the classifications described in Section 4.1 and the clustering results explained in Section 5.1, according to the following convention:

J1-J4: human judge annotations (classes: I, IQ, Q, QR, R).
CC: final consensus classification (classes: I, IQ, Q, QR, R).
s3: semantic solution in 3 clusters (labels: *clusters* 0, 1, 2).
s5: semantic solution in 5 clusters (labels: *clusters* 0, 1, 2, 3, 4).
d3: distributional solution in 3 clusters (labels: *clusters* 0, 1, 2).
d5: distributional solution in 5 clusters (labels: *clusters* 0, 1, 2, 3, 4).

Lemma	**J1**	**J2**	**J3**	**J4**	**CC**	**s3**	**s5**	**d3**	**d5**
accidental	Q	Q	Q	R	Q	1	0	1	2
accidentat	Q	Q	IQ	QR	Q	1	3	1	2
adquisitiu	R	R	R	R	R	0	1	0	1
alemany	QR	QR	R	QR	QR	1	0	0	0
alfabètic	R	R	R	R	R	1	0	0	0
alienant	Q	Q	Q	Q	Q	2	2	1	2
antic	IQ	IQ	IQ	IQ	IQ	2	2	1	2
anticlerical	Q	Q	QR	Q	Q	1	0	0	0
avergonyit	Q	Q	Q	Q	Q	1	3	1	2
bastard	Q	Q	Q	Q	Q	0	1	2	4
benigne	Q	Q	Q	Q	Q	2	2	1	2
caracurt	R	Q	Q	Q	Q	1	0	0	0
carbònic	R	R	R	R	R	0	1	2	4
celest	QR	Q	Q	R	QR	1	0	0	0
cervical	R	R	R	R	R	0	1	1	2
climatològic	R	R	R	R	R	0	1	0	0
coherent	Q	Q	Q	Q	Q	2	2	1	2
colpidor	Q	Q	Q	Q	Q	2	2	1	2
col·laborador	QR	R	R	Q	R	0	1	0	1
contaminant	Q	QR	R	Q	QR	2	2	1	2
contradictori	Q	Q	Q	Q	Q	2	2	1	2
cosmopolita	Q	Q	Q	Q	Q	2	2	1	2
cultural	QR	QR	QR	QR	QR	0	1	2	3
curatiu	Q	R	R	QR	R	0	1	2	3
destructor	Q	QR	R	QR	Q	1	3	0	1

Lemma	**J1**	**J2**	**J3**	**J4**	**CC**	**s3**	**s5**	**d3**	**d5**
diofàntic	R	R	R	R	R	0	1	0	0
diversificador	Q	Q	Q	Q	Q	1	0	2	4
duratiu	Q	Q	R	Q	Q	0	1	2	3
escàpol	Q	Q	Q	Q	Q	2	2	1	2
esfereïdor	Q	Q	Q	Q	Q	2	2	0	1
evident	Q	Q	Q	Q	Q	2	2	1	2
exempt	Q	Q	Q	Q	Q	2	2	1	2
expeditiu	Q	Q	Q	Q	Q	2	2	1	2
femení	Q	QR	QR	QR	QR	1	0	0	0
formatiu	Q	R	R	Q	R	0	1	2	3
fortuït	Q	Q	Q	Q	Q	1	3	1	2
freudià	QR	R	R	QR	R	0	1	0	0
governatiu	R	R	R	R	R	0	1	0	0
gradual	Q	Q	Q	Q	Q	1	0	2	3
grandiós	Q	Q	Q	Q	Q	2	2	1	2
gratuït	Q	Q	Q	Q	Q	2	2	1	2
honest	Q	Q	Q	Q	Q	2	2	1	2
implacable	Q	Q	Q	Q	Q	2	2	1	2
indicador	Q	QR	R	QR	R	1	0	0	1
infreqüent	Q	Q	Q	Q	Q	2	2	1	2
innoble	Q	Q	Q	Q	Q	2	2	1	2
inquiet	Q	Q	Q	Q	Q	2	2	1	2
insalvable	Q	Q	Q	Q	Q	2	2	1	2
inservible	Q	Q	Q	Q	Q	2	2	1	2
invers	Q	QR	Q	Q	Q	1	4	2	4
irreductible	Q	Q	Q	Q	Q	2	2	1	2
irònic	Q	QR	Q	QR	QR	2	2	1	2
laberíntic	Q	Q	R	Q	Q	2	2	1	2
llaminer	Q	Q	Q	Q	Q	2	2	1	2
malalt	Q	Q	Q	Q	Q	2	2	1	2
menorquí	R	QR	R	QR	QR	1	0	0	0
militar	QR	QR	R	QR	QR	0	1	0	0
morat	Q	Q	Q	Q	Q	1	4	1	2
negatiu	Q	Q	Q	QR	Q	1	0	0	1
nombrós	Q	Q	IQ	Q	Q	2	2	1	2
onomàstic	R	R	R	R	R	0	1	0	1
parlant	Q	R	R	R	R	1	0	1	2
penitenciari	R	R	R	R	R	0	1	2	4
penós	Q	Q	IQ	Q	Q	2	2	1	2
periglacial	R	R	R	R	R	0	1	2	3
pesquer	R	R	R	R	R	0	1	0	0
petri	Q	Q	R	R	R	2	2	0	0
preeminent	Q	Q	Q	Q	Q	2	2	0	1
preescolar	R	R	R	R	R	1	0	1	2
preponderant	Q	Q	Q	Q	Q	0	1	0	1

Lemma	**J1**	**J2**	**J3**	**J4**	**CC**	**s3**	**s5**	**d3**	**d5**
protector	Q	Q	QR	QR	R	1	0	1	2
raonable	Q	Q	Q	Q	Q	2	2	1	2
real	Q	QR	Q	Q	Q	2	2	0	1
representatiu	Q	Q	QR	Q	Q	2	2	1	2
salvador	Q	R	Q	QR	R	1	0	0	1
sobrenatural	Q	QR	Q	QR	Q	1	0	0	0
sociocultural	R	R	R	R	R	0	1	0	1
sonor	Q	Q	QR	QR	QR	0	1	0	0
subsidiari	Q	R	Q	Q	Q	0	1	0	1
sud-africà	R	R	R	QR	R	1	0	0	1
supraracional	Q	R	Q	R	R	0	1	2	3
terciari	R	R	R	R	R	0	1	2	3
terminològic	R	R	R	R	R	1	4	0	0
topogràfic	R	R	R	R	R	0	1	0	1
toràcic	R	R	R	R	R	0	1	0	0
triomfal	Q	QR	R	QR	QR	1	0	0	1
trivial	Q	Q	Q	Q	Q	2	2	1	2
tàctil	R	R	R	Q	R	1	0	2	3
uniforme	Q	Q	Q	Q	Q	2	2	1	2
usual	Q	Q	Q	Q	Q	2	2	1	2
utòpic	Q	Q	Q	QR	Q	2	2	1	2
vaginal	R	R	R	R	R	0	1	2	3
valencianoparlant	R	R	R	R	R	1	0	1	2
ventral	R	R	R	R	R	1	0	0	1
veterinari	R	R	R	R	R	1	0	2	4
viril	Q	Q	QR	QR	QR	2	2	2	4
vitalista	Q	QR	Q	Q	Q	2	2	1	2
vocàlic	R	R	R	R	R	0	1	0	0
xinès	R	R	R	QR	R	1	0	0	0

C.2 Gold standard B

Gold standard B consists of the 80 lemmata listed below. The list includes the classifications described in Section 4.1 and the clustering results explained in Section 5.2, according to the following convention:

J1-J3: human judge annotations (categories: B, E, O).
M: derivational morphology (categories: N, O, P, V).
Bin: binary (y = yes, n = no).
CC: consensus classfication (categories: B, E, O).
CS: clustering solution, according to their equivalence to classes (categories: B, E, O).

Lemma	J1	J2	J3	M	Bin	CC	CS
angoixós	B	E	B	N	n	B	B
artesanal	O	O	B	N	n	B	O
associat	E	E	B	P	y	E	E
atent	B	B	B	N	y	B	E
atmosfèric	O	O	O	N	n	O	O
autonòmic	O	O	O	N	n	O	O
catòlic	O	O	B	O	n	O	O
clandestí	B	B	B	O	n	B	O
clavat	E	E	E	P	y	E	E
clos	E	E	E	P	n	E	E
cognitiu	O	O	O	V	n	O	O
cordial	B	B	B	O	n	B	B
corporal	O	O	O	N	n	O	O
dedicat	E	B	E	P	y	E	E
desinteressat	B	B	B	P	n	B	B
determinant	E	E	E	V	y	E	E
eclesial	O	O	O	N	n	O	O
elemental	B	B	B	N	n	B	B
empresarial	O	O	O	N	n	O	O
episcopal	O	O	O	N	n	O	O
esmorteït	E	E	E	P	n	E	B
esperançat	B	E	B	P	n	B	B
estètic	O	O	O	N	n	O	O
eucarístic	O	O	O	N	n	O	O
explícit	B	B	B	O	n	B	B
feixuc	B	B	B	O	n	B	B
filològic	O	O	O	N	n	O	O
fix	B	B	B	O	n	B	O
foradat	E	E	E	P	n	E	O
funerari	O	O	O	N	n	O	O
gramatical	O	O	O	N	n	O	O
grisenc	B	B	B	O	n	B	B
hàbil	B	B	B	O	n	B	B
horrible	B	B	B	V	n	B	B
ignorant	B	B	E	V	n	B	B

Lemma	**J1**	**J2**	**J3**	**M**	**Bin**	**CC**	**CS**
indoeuropeu	O	O	O	N	n	O	O
inestable	B	B	B	V	n	B	B
influent	E	E	E	V	y	E	B
innat	B	B	B	P	n	B	O
interminable	E	E	E	V	n	E	B
jove	B	B	B	O	n	B	B
luxós	B	B	B	N	n	B	B
mecànic	B	O	B	N	n	B	O
mensual	B	O	O	N	n	O	O
migrat	B	B	B	P	n	B	B
militar	B	B	O	N	n	O	O
moderat	B	E	B	P	n	B	B
modernista	O	O	O	N	n	O	O
naval	O	O	O	N	n	O	O
notarial	O	O	O	N	n	O	O
obligatori	B	B	B	V	n	B	E
ocorregut	E	E	E	P	y	E	E
ofensiu	E	B	E	V	n	E	B
ornamental	B	O	B	N	n	B	O
paradoxal	B	B	B	N	n	B	B
patriòtic	B	O	B	N	n	B	O
penetrant	E	E	E	V	n	E	B
poderós	B	B	B	N	n	B	B
polit	E	E	E	P	n	E	B
problemàtic	B	B	B	N	n	B	B
radial	O	O	O	N	n	O	O
radiant	B	E	B	V	n	B	E
radioactiu	O	B	B	N	n	B	O
radiofònic	O	O	O	N	n	O	O
raonable	B	B	B	V	n	B	B
rebel	B	B	B	O	n	B	E
rectilini	B	B	B	O	n	B	O
relaxat	E	E	B	P	n	E	B
representat	E	E	E	P	y	E	E
repressiu	E	O	O	V	n	E	O
ros	B	B	B	O	n	B	B
sa	B	B	B	O	n	B	B
sever	B	B	B	O	n	B	B
singular	B	O	B	N	n	B	B
sociocultural	O	O	O	N	n	O	O
temàtic	O	O	O	N	n	O	O
tèrbol	B	B	B	O	n	B	B
terciari	B	O	B	O	n	O	O
titular	O	B	B	N	n	O	O
voluminós	B	B	B	N	n	B	B

C.3 Gold standard C

Gold standard C consists of 210 lemmata. For space reasons, we split the information associated to each lemma in two lists. Section C.3.1 contains the data concerning the assessment of agreement as explained in Chapter 4. Section C.3.2 contains the classifications obtained with each level of description in the machine learning experiments reported in Chapter 6. The information contained in each column is specified in each section.

C.3.1 Data from web experiment

M: derivational morphology (categories: N, O, P, V).
Suff: derivational suffix.
Freq: frequency of the adjective in the CTILC corpus.
#J: number of judgements obtained.
B, . . . , O: number of assignments to each class.
Entr: Entropy.
Exp: expert classifications (classes: B, BE, BO, E, EO, E).

Lemma	M	Suff	Freq	#J	B	BE	BO	E	EO	O	Entr	Exp
absort	O	-	51	49	44	0	4	0	0	1	0.38	B
abundant	V	nt	639	39	13	4	9	6	4	3	1.65	E
abundós	V	ós	228	56	12	10	5	17	4	8	1.68	E
acompanyat	P	t	848	39	16	2	10	5	3	3	1.52	E
admirable	V	ble	309	39	3	2	1	21	9	3	1.31	E
aleatori	V	ori	72	34	21	1	3	2	1	6	1.19	B
alegre	O	-	388	42	17	1	20	0	0	4	1.03	BO
altiu	O	-	74	43	24	1	7	1	0	10	1.13	B
americà	N	à	622	43	0	0	1	0	2	40	0.29	O
amorós	N	ós	410	38	1	1	6	2	5	23	1.20	BO
ample	O	-	1191	58	33	1	21	3	0	0	0.91	B
anarquista	N	ista	189	48	2	0	5	0	1	40	0.60	BO
angular	N	ar	54	42	4	0	2	1	0	35	0.60	O
animal	O	-	434	40	20	1	11	0	1	7	1.19	B
animat	P	t	248	36	9	6	8	8	2	3	1.68	BE
anòmal	O	-	73	39	19	1	14	0	0	5	1.07	B
atòmic	N	ic	189	42	1	0	2	1	4	34	0.71	O
baix	O	-	2691	41	23	3	13	0	1	1	1.06	B
barceloní	N	í	679	46	1	0	3	0	0	42	0.34	O
benigne	O	-	179	42	25	1	14	0	0	2	0.90	B
bord	O	-	131	51	41	2	2	4	0	2	0.75	B
caduc	O	-	71	43	16	9	3	10	2	3	1.54	B
calb	O	-	69	29	8	0	2	0	0	19	0.81	B
calcari	N	ari	163	47	1	0	1	3	4	38	0.72	O
capaç	O	-	1979	36	2	4	14	6	1	9	1.51	B
capitalista	N	ista	663	58	5	0	16	2	1	34	1.06	BO
cardinal	N	al	73	35	19	0	6	0	0	10	0.99	B
catalanista	N	ista	152	42	2	0	7	2	1	30	0.91	BO
causal	N	al	203	39	4	0	1	2	4	28	0.95	O
caut	O	-	55	40	22	0	8	2	1	7	1.19	B

Lemma	M	Suff	Freq	#J	B	BE	BO	E	EO	O	Entr	Exp
cèlebre	O	-	289	44	27	0	8	0	0	9	0.93	B
ciutadà	N	à	273	41	1	0	4	0	1	35	0.54	O
comptable	V	ble	81	38	1	3	0	14	16	4	1.26	EO
comunista	N	ista	487	56	7	0	13	0	0	36	0.88	BO
concret	O	-	2331	40	27	7	1	2	0	3	1.00	B
conflictiu	N	iu	191	41	7	0	13	1	1	19	1.20	O
conservador	V	or	329	46	23	15	1	5	2	0	1.17	B
contingent	O	-	91	32	11	4	2	3	0	12	1.39	B
contradictori	V	ori	371	41	9	2	0	22	2	6	1.24	E
convincent	V	nt	109	43	3	2	0	33	4	1	0.84	E
cooperatiu	V	iu	115	39	1	1	0	14	16	7	1.22	EO
corporatiu	V	iu	57	38	1	0	2	1	0	34	0.44	O
cranià	N	à	81	40	0	0	0	0	0	40	0	O
creador	V	or	403	42	0	4	2	20	14	2	1.23	E
cridaner	V	er	80	41	2	10	0	20	8	1	1.25	BE
cru	O	-	240	40	25	4	7	2	1	1	1.16	B
curull	O	-	79	33	26	1	1	5	0	0	0.68	B
decisiu	V	iu	700	42	7	2	6	15	4	8	1.62	B
deficient	O	-	147	41	20	3	10	1	0	7	1.27	B
deliciós	N	ós	154	38	19	2	8	2	0	7	1.29	B
desproporcionat	P	t	65	38	11	1	11	1	1	13	1.37	B
diari	N	ari	519	36	4	1	4	0	0	27	0.80	O
dificultós	V	ós	78	45	14	3	16	1	1	10	1.41	B
digestiu	V	iu	140	56	1	2	4	13	16	20	1.44	EO
diürn	N	altres	74	39	10	0	16	0	0	13	1.08	BO
divergent	V	nt	72	39	7	11	1	13	5	2	1.54	E
docent	V	nt	101	55	0	0	3	3	12	37	0.91	EO
elèctric	N	ic	719	42	1	0	0	1	5	35	0.58	O
embolicat	P	t	165	39	6	11	1	15	4	2	1.49	BE
encantat	P	t	105	58	17	11	8	16	1	5	1.58	BE
encarregat	P	t	273	34	5	4	1	17	0	7	1.30	E
epistemològic	N	ic	89	31	0	0	1	0	1	29	0.28	O
eròtic	N	ic	163	40	2	2	7	1	5	23	1.27	BO
escènic	N	ic	179	42	1	0	3	3	9	26	1.09	O
esquerre	O	-	577	41	34	0	7	0	0	0	0.45	B
estacional	N	al	81	48	5	0	9	0	2	32	0.95	O
excels	O	-	62	33	16	2	3	7	0	5	1.35	B
exigent	V	nt	168	43	4	3	0	25	7	4	1.23	E
exportador	V	or	126	37	1	5	0	20	9	2	1.20	E
exquisit	O	-	125	41	22	2	7	1	1	8	1.28	B
familiar	N	ar	1271	41	2	0	10	0	2	27	0.91	BO
fangós	N	ós	53	37	1	0	1	1	4	30	0.70	O
feminista	N	ista	181	43	16	1	11	0	0	15	1.17	BO
fluix	O	-	199	44	31	2	6	2	2	1	1.02	B
foll	O	-	132	48	32	0	10	3	0	3	0.94	B

Lemma	M	Suff	Freq	#J	B	BE	BO	E	EO	O	Entr	Exp
formidable	O	-	120	41	40	0	1	0	0	0	0.11	B
franc	O	-	275	46	23	1	15	0	0	7	1.08	B
fresc	O	-	709	47	20	4	11	1	2	9	1.44	B
gros	O	-	1666	41	28	3	8	0	0	2	0.91	B
gruixut	O	-	374	42	26	0	13	0	0	3	0.84	B
humà	N	à	5166	55	7	0	17	4	0	27	1.16	BO
humil	O	-	415	45	21	0	14	0	0	10	1.05	B
igual	O	-	1725	40	27	1	10	1	0	1	0.88	B
immutable	V	ble	154	38	3	14	3	14	3	1	1.43	E
imperceptible	V	ble	85	37	2	8	0	20	6	1	1.21	E
imperfecte	O	-	91	41	23	3	13	1	0	1	1.06	B
imperial	N	al	142	47	1	0	6	2	0	38	0.65	O
impropi	O	-	51	45	38	0	7	0	0	0	0.43	B
incomplet	O	-	173	42	27	12	1	2	0	0	0.87	B
infantil	O	-	655	46	8	0	22	0	0	16	1.02	BO
informatiu	V	iu	311	40	2	0	2	14	17	5	1.29	E
inhumà	N	à	67	47	34	2	6	1	1	3	0.97	B
insuficient	V	nt	288	42	29	2	8	1	0	2	0.95	B
integral	O	-	168	39	16	2	3	9	4	5	1.55	B
íntegre	O	-	96	55	25	6	9	5	0	10	1.42	B
intel·ligent	V	nt	426	40	9	0	8	0	6	17	1.30	B
intern	O	-	1339	55	37	2	12	2	0	2	0.96	B
intuïtiu	V	iu	101	42	0	2	1	9	13	17	1.29	BO
irat	P	t	54	43	12	2	11	0	1	17	1.30	E
líquid	O	-	186	39	26	3	5	3	0	2	1.08	B
llarg	O	-	4701	48	34	4	9	0	0	1	0.84	B
lleidatà	N	à	134	37	0	0	1	0	0	36	0.12	O
llis	O	-	323	49	33	10	1	5	0	0	0.90	B
local	N	al	1907	41	12	0	7	0	2	20	1.15	BO
mal	O	-	2382	21	5	0	3	0	1	12	1.08	B
manresà	N	à	71	41	0	0	1	0	0	40	0.11	O
marxià	N	ià	73	33	0	0	4	0	0	29	0.36	O
matiner	V	er	53	45	3	3	0	26	8	5	1.22	E
màxim	O	-	1272	36	29	4	0	2	0	1	0.67	B
melòdic	N	ic	103	41	0	0	10	0	1	30	0.66	O
menor	O	-	1118	42	38	0	4	0	0	0	0.31	B
mercantil	O	-	209	38	3	0	2	0	1	32	0.59	O
mínim	O	-	857	57	40	12	2	1	0	2	0.88	B
moll	O	-	133	44	17	16	2	5	0	4	1.34	B
morat	N	at	172	29	8	0	1	0	2	18	0.95	B
motor	V	or	145	35	3	0	2	13	3	14	1.31	E
mutu	O	-	355	35	31	0	0	0	0	4	0.35	B
nocturn	N	altres	369	48	11	0	27	0	0	10	0.98	BO
notori	V	ori	163	44	22	4	7	6	0	5	1.37	B
nutritiu	V	iu	153	37	1	2	0	12	12	10	1.33	EO

Lemma	M	Suff	Freq	#J	B	BE	BO	E	EO	O	Entr	Exp
obert	P	t	2250	38	15	14	1	7	1	0	1.23	BE
oblidat	P	t	337	42	8	10	5	12	4	3	1.68	E
obrer	V	er	923	53	11	3	5	7	4	23	1.53	O
ocult	O	-	195	45	19	15	1	10	0	0	1.14	B
ontològic	N	ic	89	41	1	0	2	0	0	38	0.30	O
opac	O	-	153	43	29	1	10	0	0	3	0.87	B
orientat	P	t	285	40	7	7	5	11	4	6	1.73	E
paradoxal	N	al	131	34	6	0	5	1	0	22	0.97	B
pasqual	N	al	82	38	0	0	0	0	0	38	0	O
peculiar	N	ar	317	42	26	0	10	1	0	5	0.98	B
peninsular	N	ar	219	40	1	0	3	0	0	36	0.38	O
perillós	N	ós	808	41	9	0	10	0	2	20	1.17	B
pertinent	V	nt	135	41	17	4	7	8	0	5	1.46	B
pessimista	N	ista	83	57	41	1	12	0	0	3	0.79	B
picat	P	t	260	40	4	9	2	22	2	1	1.28	E
plàcid	O	-	103	45	21	0	12	1	1	10	1.21	B
poètic	N	ic	519	39	2	0	2	0	1	34	0.51	BO
precoç	O	-	148	36	19	0	6	0	0	11	0.99	B
predilecte	O	-	59	39	22	0	11	0	0	6	0.96	B
preferible	V	ble	104	38	4	2	2	16	2	12	1.43	E
primari	N	ari	681	42	35	1	3	2	0	1	0.66	B
primitiu	V	iu	1053	41	28	0	3	0	0	10	0.79	B
productor	V	or	200	42	0	2	1	28	9	2	0.97	E
professional	N	al	978	39	10	2	7	3	2	15	1.52	BO
promès	P	t	90	37	0	0	0	11	12	14	1.09	E
propens	O	-	89	30	12	1	2	5	0	10	1.32	B
pròsper	O	-	66	38	8	2	1	5	7	15	1.52	B
protector	V	or	145	38	0	2	0	24	11	1	0.89	E
prudent	O	-	236	54	19	1	17	0	0	17	1.16	B
punxegut	N	ut	74	34	1	3	2	17	3	8	1.38	B
quadrat	P	t	412	49	17	3	3	6	1	19	1.41	B
raonable	V	ble	249	41	2	2	4	19	7	7	1.48	BE
reaccionari	N	ari	135	42	18	2	4	9	6	3	1.52	B
recent	O	-	809	37	29	0	3	0	0	5	0.66	B
receptor	V	or	57	42	6	9	0	19	8	0	1.28	E
recíproc	O	-	206	36	10	0	9	0	1	16	1.16	B
recomanat	P	t	86	37	5	2	0	20	4	6	1.29	E
regulador	V	or	93	38	0	1	0	28	8	1	0.74	E
remarcable	V	ble	323	41	11	11	0	19	0	0	1.06	B
renaixentista	N	ista	100	57	0	0	3	1	2	51	0.44	O
respiratori	V	ori	161	54	0	0	1	7	5	41	0.76	O
responsable	V	ble	530	57	24	0	9	3	2	19	1.29	B
resultant	V	nt	153	40	3	3	0	19	8	7	1.36	E
revelador	V	or	102	56	7	4	0	38	5	2	1.04	E
revolucionari	N	ari	772	43	5	2	6	2	14	14	1.54	BO

Lemma	M	Suff	Freq	#J	B	BE	BO	E	EO	O	Entr	Exp
rígid	O	-	309	42	15	8	7	5	4	3	1.64	B
roent	V	nt	58	40	11	6	3	14	1	5	1.55	B
sabut	P	t	430	41	6	11	1	14	6	3	1.56	BE
salvador	V	or	74	42	0	0	1	26	13	2	0.89	E
sant	O	-	1038	34	3	1	3	9	7	11	1.57	B
satisfactori	V	ori	233	38	5	7	0	16	6	4	1.47	E
semicircular	N	ar	67	31	4	0	1	0	0	26	0.52	B
sensitiu	V	iu	70	41	3	1	1	6	8	22	1.30	BO
seriós	N	ós	735	39	10	0	5	0	1	23	1.01	B
significatiu	V	iu	832	46	30	3	3	2	1	7	1.14	B
silenciós	N	ós	303	55	17	6	20	0	1	11	1.36	B
similar	N	ar	618	40	24	3	8	0	0	5	1.08	B
simplista	N	ista	53	43	12	0	6	18	2	5	1.38	B
socialista	N	ista	811	41	3	1	6	3	1	27	1.12	BO
sospitós	V	ós	144	40	7	3	3	6	3	18	1.53	E
subaltern	O	-	67	46	27	1	2	8	0	8	1.14	B
sublim	O	-	100	35	22	3	0	4	1	5	1.12	B
subsidiari	N	ari	51	34	15	1	5	1	1	11	1.31	B
subterrani	O	-	211	27	5	1	2	6	3	10	1.57	B
superflu	O	-	82	43	36	3	1	2	0	1	0.65	B
temible	V	ble	92	43	1	1	2	20	12	7	1.32	E
tenaç	O	-	71	56	26	1	10	1	0	18	1.17	B
terrestre	N	altres	535	42	4	0	9	0	1	28	0.91	O
terrible	V	ble	561	41	17	2	13	1	2	6	1.39	B
típic	N	ic	990	57	43	1	7	2	2	2	0.89	B
titular	N	ar	353	49	10	0	7	2	2	28	1.18	B
tort	O	-	57	42	18	14	2	8	0	0	1.19	B
total	N	al	2262	45	21	3	5	4	0	12	1.34	B
tou	O	-	259	45	32	5	7	0	0	1	0.86	B
treballador	V	or	250	38	4	4	1	18	7	4	1.47	E
triangular	N	ar	105	38	2	0	0	1	1	34	0.44	B
turístic	N	ístic	287	39	0	0	2	0	3	34	0.46	BO
unitari	N	ari	316	45	5	1	6	7	3	23	1.41	BO
utilitari	N	ari	84	48	5	1	4	15	4	19	1.46	BO
vague	O	-	228	33	19	2	6	2	1	3	1.29	B
variable	V	ble	428	41	5	14	3	16	3	0	1.37	E
vegetatiu	V	iu	154	35	4	2	0	14	5	10	1.41	EO
ver	O	-	475	47	21	0	14	2	1	9	1.25	B
viari	N	ari	56	47	1	0	2	1	0	43	0.37	O
viciós	N	ós	73	39	3	0	6	3	5	22	1.26	B
victoriós	N	ós	84	40	7	1	13	5	3	11	1.57	E
vigorós	N	ós	134	40	13	2	8	0	0	17	1.20	B
viril	O	-	56	39	9	0	13	0	0	17	1.06	B
vivent	V	nt	505	46	8	7	1	21	6	3	1.47	E
vulgar	N	ar	319	42	23	3	9	0	1	6	1.21	B

C.3.2 Data from machine learning experiments

Exp: expert classifications.
Morph: classification with morphological features.
Func: classification with features related to syntactic function.
Uni: classification with unigram features.
Bi: classification with bigram features.
Sem: classification with semantic features.
All: classification with all features.

All classifications distinguish between the following classes: B, BE, BO, E, EO, E.

Lemma	Exp	Morph	Func	Uni	Bi	Sem	All
absort	B	B	B	B	B	B	B
abundant	E	E	B	B	B	B	BE
abundós	E	B	B	B	B	B	BE
acompanyat	E	BE	B	B	B	B	BE
admirable	E	E	B	B	B	B	B
aleatori	B	E	B	B	B	B	BE
alegre	BO	B	B	B	B	B	B
altiu	B	B	B	B	B	B	B
americà	O	BO	B	O	B	B	BO
amorós	BO	B	B	B	B	B	B
ample	B	B	B	B	B	B	B
anarquista	BO	BO	B	B	B	B	B
angular	O	B	BO	B	BE	B	B
animal	B	B	BO	BO	BO	BO	B
animat	BE	BE	B	B	B	B	BE
anòmal	B	B	B	B	B	B	B
atòmic	O	BO	B	BO	BO	B	O
baix	B	B	B	B	B	B	B
barceloní	O	B	BO	B	B	BO	B
benigne	B	B	B	B	B	B	B
bord	B	B	BO	B	B	B	B
caduc	B	B	B	B	BO	B	B
calb	B	B	B	B	B	B	B
calcari	O	BO	B	BO	BO	B	B
capaç	B	B	B	B	B	B	B
capitalista	BO	BO	B	BO	BO	B	BO
cardinal	B	BO	B	B	B	B	B
catalanista	BO	BO	B	B	B	B	BO
causal	O	BO	BO	O	O	B	BO
caut	B	B	B	B	B	B	B

Lemma	Exp	Morph	Func	Uni	Bi	Sem	All
cèlebre	B	B	B	B	B	B	B
ciutadà	O	BO	O	B	B	O	BO
comptable	EO	E	O	O	B	B	EO
comunista	BO	BO	B	B	B	B	B
concret	B	B	B	B	B	B	B
conflictiu	O	BO	B	B	B	B	B
conservador	B	E	B	B	B	B	BE
contingent	B	B	B	B	B	B	B
contradictori	E	E	B	B	B	B	BE
convincent	E	E	B	B	B	B	BE
cooperatiu	EO	E	O	O	O	BO	EO
corporatiu	O	E	B	B	B	B	E
cranià	O	BO	O	BO	B	B	O
creador	E	E	B	B	B	B	BE
cridaner	BE	E	B	B	B	B	BE
cru	B	B	B	B	B	B	B
curull	B	B	B	B	B	B	B
decisiu	B	E	B	B	B	B	BE
deficient	B	B	B	B	B	B	B
deliciós	B	B	B	B	B	B	B
desproporcionat	B	BE	B	B	B	B	BE
diari	O	BO	BO	BO	O	B	B
dificultós	B	E	B	B	B	B	BE
digestiu	EO	E	O	BO	BO	EO	EO
diürn	BO	BO	BO	B	B	B	B
divergent	E	E	B	B	B	B	BE
docent	EO	E	O	BO	BO	O	BE
elèctric	O	BO	BO	BO	BO	B	B
embolicat	BE	E	B	B	B	B	BE
encantat	BE	E	B	B	B	B	BE
encarregat	E	BE	B	B	B	B	BE
epistemològic	O	BO	O	B	B	B	B
eròtic	BO	BO	B	B	B	B	B
escènic	O	BO	O	BO	O	O	O
esquerre	B	B	O	BO	E	O	B
estacional	O	BO	B	B	B	B	B
excels	B	B	B	B	B	B	B
exigent	E	E	B	B	B	B	BE
exportador	E	E	B	B	BE	B	E
exquisit	B	B	B	B	B	B	B
familiar	BO	B	B	B	B	B	B
fangós	O	B	B	B	B	B	B
feminista	BO	BO	B	B	B	B	B
fluix	B	B	B	B	B	B	B
foll	B	B	B	B	B	B	B

Lemma	**Exp**	**Morph**	**Func**	**Uni**	**Bi**	**Sem**	**All**
formidable	B	B	B	B	B	B	B
franc	B	B	B	B	B	B	B
fresc	B	B	B	B	B	B	B
gros	B	B	B	B	B	B	B
gruixut	B	B	B	B	B	B	B
humà	BO	BO	B	B	BO	B	B
humil	B	B	B	B	B	B	B
igual	B	B	B	B	B	B	B
immutable	E	E	BE	B	B	B	BE
imperceptible	E	E	B	B	B	B	BE
imperfecte	B	B	B	B	B	B	B
imperial	O	BO	BO	B	B	BO	BO
impropi	B	B	B	B	B	B	B
incomplet	B	B	B	B	B	B	B
infantil	BO	B	B	B	B	B	B
informatiu	E	EO	B	BO	BO	BO	EO
inhumà	B	BO	B	B	B	B	B
insuficient	B	E	B	B	B	B	B
integral	B	B	BO	B	B	B	B
íntegre	B	B	BE	B	B	B	B
intel·ligent	B	E	B	B	B	B	BE
intern	B	B	B	B	B	B	B
intuïtiu	BO	E	B	B	BO	B	BE
irat	E	BE	B	B	B	BE	B
líquid	B	B	B	B	B	B	B
llarg	B	B	B	B	B	B	B
lleidatà	O	BO	BO	O	B	B	BO
llis	B	B	B	B	B	B	B
local	BO	BO	BO	O	BO	BO	BO
mal	B	B	B	B	B	B	B
manresà	O	BO	B	B	B	B	B
marxià	O	B	O	BO	B	O	O
matiner	E	E	B	B	B	B	BE
màxim	B	B	B	B	B	B	B
melòdic	O	BO	BO	BO	BO	B	BO
menor	B	B	B	B	B	B	B
mercantil	O	B	O	BO	B	O	B
mínim	B	B	B	B	B	B	B
moll	B	B	B	B	B	B	B
morat	B	B	B	B	B	B	B
motor	E	E	O	BO	BO	O	E
mutu	B	B	B	B	B	B	B
nocturn	BO	BO	B	B	B	B	B
notori	B	B	B	B	B	B	BE
nutritiu	EO	E	B	B	BO	B	BE

Lemma	**Exp**	**Morph**	**Func**	**Uni**	**Bi**	**Sem**	**All**
obert	BE	E	B	B	B	B	BE
oblidat	E	BE	B	B	B	B	BE
obrer	O	E	BO	BO	BO	BO	BE
ocult	B	B	B	B	B	B	B
ontològic	O	BO	BO	BO	BO	O	B
opac	B	B	B	B	B	B	B
orientat	E	BE	B	B	B	B	BE
paradoxal	B	BO	B	B	B	B	B
pasqual	O	BO	O	BO	B	B	O
peculiar	B	B	B	B	B	B	B
peninsular	O	B	O	BO	BO	O	B
perillós	B	B	B	B	B	B	B
pertinent	B	E	B	B	B	B	BE
pessimista	B	BO	B	B	B	BE	B
picat	E	BE	B	B	B	B	BE
plàcid	B	B	B	B	B	B	B
poètic	BO	BO	B	BO	BO	B	B
precoç	B	B	BO	B	B	B	B
predilecte	B	B	B	B	B	B	B
preferible	E	E	B	B	B	BE	BE
primari	B	BO	B	BO	BO	B	B
primitiu	B	EO	B	B	BO	B	BE
productor	E	E	B	B	B	B	BE
professional	BO	BO	BO	BO	B	BO	BO
promès	E	E	B	B	B	B	BE
propens	B	B	B	B	B	B	B
pròsper	B	B	B	B	B	B	B
protector	E	E	BO	B	B	B	BE
prudent	B	B	B	B	B	B	B
punxegut	B	B	B	B	B	B	B
quadrat	B	BE	B	B	B	B	EO
raonable	BE	E	B	B	B	B	BE
reaccionari	B	BO	B	B	B	B	B
recent	B	B	B	B	B	B	B
receptor	E	E	B	B	B	B	E
recíproc	B	B	B	B	B	B	B
recomanat	E	BE	B	B	B	B	BE
regulador	E	E	B	B	B	B	E
remarcable	B	E	B	B	B	B	BE
renaixentista	O	BO	B	B	B	B	B
respiratori	O	E	O	BO	BO	O	E
responsable	B	E	B	B	B	B	B
resultant	E	E	BO	B	B	B	E
revelador	E	E	B	B	B	B	BE
revolucionari	BO	BO	BO	B	B	B	B

Lemma	**Exp**	**Morph**	**Func**	**Uni**	**Bi**	**Sem**	**All**
rígid	B	B	B	B	B	B	B
roent	B	E	B	B	B	B	BE
sabut	BE	E	B	B	B	B	B
salvador	E	E	B	B	B	B	E
sant	B	B	B	B	B	B	B
satisfactori	E	B	B	B	B	B	BE
semicircular	B	B	B	B	B	O	B
sensitiu	BO	E	O	B	B	B	E
seriós	B	B	B	B	B	B	B
significatiu	B	EO	B	B	B	B	B
silenciós	B	B	B	B	B	B	B
similar	B	B	B	B	B	B	B
simplista	B	BO	B	B	B	B	B
socialista	BO	BO	B	B	B	BO	B
sospitós	E	B	B	B	B	B	BE
subaltern	B	B	B	B	B	B	B
sublim	B	B	B	B	B	B	B
subsidiari	B	BO	BO	B	B	B	B
subterrani	B	B	B	B	B	B	B
superflu	B	B	B	B	B	B	B
temible	E	E	B	B	B	B	B
tenaç	B	B	B	B	B	B	B
terrestre	O	BO	B	B	BO	B	B
terrible	B	E	B	B	B	B	BE
típic	B	BO	B	B	B	B	B
titular	B	B	B	B	B	B	B
tort	B	B	E	B	B	B	B
total	B	BO	B	B	B	B	B
tou	B	B	B	B	B	B	B
treballador	E	E	BO	B	BO	B	B
triangular	B	B	B	B	B	B	B
turístic	BO	B	O	BO	B	BO	BO
unitari	BO	BO	B	B	B	B	B
utilitari	BO	BO	B	BO	B	B	B
vague	B	B	B	B	B	B	B
variable	E	E	B	B	B	B	E
vegetatiu	EO	E	O	B	O	B	EO
ver	B	B	B	B	B	B	B
viari	O	BO	O	BO	B	O	B
viciós	B	B	B	B	B	B	B
victoriós	E	B	B	B	B	B	B
vigorós	B	B	B	B	B	B	B
viril	B	B	B	B	B	B	B
vivent	E	E	BO	B	BO	B	BO
vulgar	B	B	B	B	B	B	B

References

Abe, N. and Li, H. (1996). Learning word association norms using tree cut pair models. In *Proceedings of the 13th International Conference on Machine Learning, ICML*, pages 3–11.

Alcover, A. and Moll, F. (2002). *Diccionari català-valencià-balear*. Barcelona: l'Institut.

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

Altaye, M., Donner, A., and Eliasziw, M. (2001). A general goodness-of-fit approach for inference procedures concerning the kappa statistic. *Statistics in Medicine*, *20*(16), 2479–2488.

Artstein, R. and Poesio, M. (2005a). Bias decreases in proportion to the number of annotators. In Jaeger, G., Monachesi, P., Penn, G., Rogers, J., and Wintner, S. (Eds.), *Proceedings of FG-MoL 2005*, pages 141–150, Edinburgh.

Artstein, R. and Poesio, M. (2005b). Kappa3 = alpha (or beta). Technical report, University of Essex Department of Computer Science.

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., and Padró, M. (2006). Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 48–55.

Badia, T., Boleda, G., Colominas, C., Garmendia, M., González, A., and Quixal, M. (2002). BancTrad: a web interface for integrated access to parallel annotated corpora. In *Proceedings of the Workshop on Language Resources for Translation Work and Research held during the 3rd LREC Conference*, Las Palmas.

Badia i Margarit, A. M. (1995). *Gramàtica de la llengua catalana descriptiva, normativa, diatòpica, diastràtica*. Barcelona: Proa.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, Montreal, Canada.

Baldewein, U., Erk, K., Pado, S., and Prescher, D. (2004). Semantic role labelling for chunk sequences. In *Proceedings of the CoNLL'04 shared task*, Boston, MA.

Bally, C. (1944). *Linguistique générale et linguistique française*. Berne: A. Francke.

Bennet, E. M., Alpert, R., and Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, *18*, 303–308.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice*. Cambridge: The MIT Press.

Bohnet, B., Klatt, S., and Wanner, L. (2002). An approach to automatic annotation of functional information to adjectives with an application to German. In *Proceedings of the 3rd LREC Conference, Workshop: Linguistic Knowledge Acquisition and Representation*.

Boleda, G. (2003). Adquisició de classes adjectivals. Master's thesis, Universitat Pompeu Fabra, Barcelona.

Boleda, G. and Alonso, L. (2003). Clustering adjectives for class acquisition. In *Proceedings of the EACL'03 Student Session*, pages 9–16, Budapest, Hungary.

Boleda, G., Badia, T., and Batlle, E. (2004). Acquisition of semantic classes for adjectives from distributional evidence. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1119–1125, Geneva, Switzerland.

Boleda, G., Badia, T., and Schulte im Walde, S. (2005). Morphology vs. syntax in adjective class acquisition. In *Proceedings of the ACL-SIGLEX 2005 Workshop on Deep Lexical Acquisition*, pages 1119–1125, Ann Arbor, USA.

Boleda, G., Bott, S., Castillo, C., Meza, R., Badia, T., and López, V. (2006). Cucweb: a catalan corpus built from the web. In Kilgarriff, A. and Baroni, M. (Eds.), *2nd Web as Corpus Workshop at EACL'06*.

Bonet, S. and Solà, J. (1986). *Sintaxi generativa catalana*. Barcelona: Enciclopèdia Catalana.

Bosque, I. and Picallo, C. (1996). Postnominal adjectives in Spanish DPs. *Journal of Linguistics*, *32*, 349–386.

Bouckaert, R. R. (2004). Estimating replicability of classifier learning experiments. In *Proceedings of the 21st International Conference on Machine Learning (ICML 2004)*, Banff, Alberta, Canada.

Bouillon, P. (1997). *Polymorphie et sémantique lexicale: le cas des adjectifs*. PhD thesis, Université de Paris 7 Denis Diderot.

Bouillon, P. (1999). The adjective "vieux": The point of view of "generative lexicon". In Viegas, E. (Ed.), *Breadth and depth of semantics lexicons*, pages 148–166. Dordrecht: Kluwer.

Brants, T. (2000). Inter-annotator agreement for a german newspaper corpus. In *Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–23.

Bresnan, J. (1982). The passive in lexical theory. In Bresnan, J. (Ed.), *The Mental Representation of Grammatical Relations*. Cambridge, Massachusetts: The MIT Press.

Bresnan, J. (1995). Lexicality and argument structure. Invited talk at the Paris Syntax and Semantics Conference.

Burnage, G. and Dunlop, D. (1992). Encoding the British National Corpus. In *Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora*.

Byrt, T., Bishop, J., and Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423–429.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, *22*(2), 249–254.

Carlson, G. (1977). *Reference to Kinds in English*. University of Massachussets: unpublished.

Carvalho, P. and Ranchhod, E. (2003). Analysis and disambiguation of nouns and adjectives in Portuguese by FST. In *Proceedings of the Workshop on Finite-State Methods for Natural Language Processing at EACL2003*, pages 105–112, Budapest, Hungary.

Chao, G. and Dyer, M. G. (2000). Word sense disambiguation of adjectives using probabilistic networks. In *Proceedings of COLING*, pages 152–158.

Chierchia, G. and McConnell Ginet, S. (2000). *Meaning and grammar an introduction to semantics* (2nd ed.). Cambridge (Mass.) [etc.]: MIT Press.

Cicchetti, D. V. and Feinsten, A. R. (1990). High agreement but low kappa: Ii. resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.

Corley, M. and Scheepers, C. (2002). Syntactic priming in english sentence production: Categorical and latency evidence from an internet-based study. *Psychonomic Bulletin and Review*, *9*(1), 126–131.

De Cuyper, G. (2006). Variaciones en el uso de ser y estar en catalán. Un panorama empírico. Submitted.

Demonte, V. (1999). El adjetivo: clases y usos. la posición del adjetivo en el sintagma nominal. In Bosque, I. and Demonte, V. (Eds.), *Gramática Descriptiva de la Lengua Española*, pages 129–215. Madrid: Espasa-Calpe.

Di Eugenio, B. and Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics*, *30*(1), 95–101.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1924.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*, 5–23.

Dietterich, T. G. (2002). Ensemble learning. In Arbib, M. A. (Ed.), *The Handbook of Brain Theory and Neural Networks* (Second ed.). Cambridge, MA: The MIT Press.

Dixon, R. M. W. (1982). Where have all the adjectives gone? In *Where Have All the Adjectives Gone? and Other Essays in Semantics and Syntax*, pages 1–62. Berlin-Amsterdam-New York: Mouton.

Dixon, R. M. W. (2004). Adjective classes in typological perspective. In Dixon, R. M. W. and Aikhenvald, A. Y. (Eds.), *Adjective Classes*, pages 1–49. Oxford: Oxford University Press.

Dixon, R. M. W. and Aikhenvald, A. Y. (Eds.). (2004). *Adjective Classes*. Oxford: Oxford University Press.

Dowty, D. R., Wall, R. E., and Peters, S. (1981). *Introduction to Montague Semantics*. Dordrecht: Reidel.

Engel, U. (1988). *Deutsche Grammatik*. Heidelberg: Julius Groos.

Everitt, B., Landau, S., and Leese, M. (2001). *Cluster Analysis* (Fourth ed.). London: Arnold.

Fellbaum, C. (1998a). A semantic network of English verbs. In Fellbaum, C. (Ed.), *WordNet: an Electronic Lexical Database*, pages 69–104. London: MIT.

Fellbaum, C. (Ed.). (1998b). *WordNet: an electronic lexical database*. London: MIT.

Fellbaum, C., Grabowski, J., and Landes, S. (1998). Performance and confidence in a semantic annotation task. In Fellbaum, C. (Ed.), *WordNet: An Electronic Lexical Database*, chapter 9, pages 217–237. Cambridge, Massachusetts: The MIT Press.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (Second ed.). Wiley series in probability and mathematical statistics. New York: John Wiley & Sons.

Fradin, B. and Kerleroux, F. (2002). Troubles with lexemes. In Geert, B., de Cesaris, J., Scalie, S., and Rallis, A. (Eds.), *Proceedings of the Third Mediterranean Meeting on Morphology*, Barcelona.

Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, *100*, 25–50.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148–156, San Francisco. Morgan Kaufmann.

Ghamrawi, N. and McCallum, A. (2005). Collective multi-label classification. In *Fourteenth Conference on Information and Knowledge Management (CIKM)*.

Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing. In Guarino, N. and Poli, R. (Eds.), *Formal Ontology in conceptual Analysis and Knowledge Representation*. Kluwer Academic Publishers.

Hamann, C. (1991). Adjectivsemantik/Adjectival Semantics. In von Stechow, A. and Wunderlich, D. (Eds.), *Semantik/Semantics. Ein internationales Handbuch der Zeitgenössischen Forschung. An International Handbook of Contemporary Research*, pages 657–673. Berlin/NY: De Gruyter.

Harris, Z. (1968). *Mathematical Structures of Language*. New York: John Wiley & Sons.

Hatzivassiloglou, V. and McKeown, K. R. (1993). Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 172–182.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of ACL/EACL*, pages 174–181.

Hatzivassiloglou, V. and McKeown, K. R. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING*, pages 299–305.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Blackwell.

Hripcsak, G. and Heitjan, D. F. (2002). Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, *35*(2), 99–110.

Huddleston, R. and Pullum, G. K. (2001). *The Cambridge Grammar of the English Language*. Cambridge (UK) [etc.]: Cambridge University Press.

Institut d'Estudis Catalans (1997). *Diccionari de la llengua catalana*. Barcelona (etc.): Edicions 67.

Jackendoff, R. S. (1990). *Semantic structures*. Cambridge (Mass.): MIT Press.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall.

Justeson, J. S. and Katz, S. M. (1995). Principled disambiguation: Discriminating adjective senses with modified nouns. *Computational Linguistics, 21*(1), 1–27.

Kamp, J. A. W. (1975). Two theories about adjectives. In Keenan, E. L. (Ed.), *Formal Semantics of Natural Language*, pages 123–155. Cambridge: Cambridge University Press.

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (Eds.). (1995). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Berlin/New York: Mouton de Gruyter.

Karypis, G. (2002). *CLUTO: A Clustering Toolkit*.

Katz, J. J. (1972). *Semantic Theory*. New York: Harper and Row.

Katz, J. J. and Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*(1), 170–210.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. New York [etc.]: John Wiley.

Keller, F., Corley, M., Corley, S., Konieczny, L., and Todirascu, A. (1998). Webexp: A java toolbox for web-based psychological experiments. Technical report, Human Communication Research Centre, University of Edinburgh.

Korhonen, A. (2002a). Semantically motivated subcategorization acquisition. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, pages 51–58, Philadelphia.

Korhonen, A. (2002b). *Subcategorization Acquisition*. PhD thesis, Computer Laboratory, University of Cambridge.

Kraemer, H. C., Periyakoil, V. S., and Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine, 21*, 2109–2129.

Krifka, M., Pelletier, F. J., Carlson, G. N., Ter meulen, A., Chierchia, G., and Link, G. (1995). Genericity: An introduction. In Carlson, G. N. and Pelletier, F. J. (Eds.), *The Generic Book*, pages 1–124. Chicago: University of Chicago Press.

Krippendorff, K. (1980). *Content Analysis : An Introduction to Its Methodology*. SAGE Publications.

Krippendorff, K. (2004a). *Content analysis : an introduction to its methodology* (Second ed.). Thousand Oaks, Calif.: Sage.

Krippendorff, K. (2004b). Reliability in content analysis. some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433.

Lahav, R. (1989). Against compositionality: The case of adjectives. *Philosophical Studies*, *57*, 261–279.

Landis, J. R. and Koch, G. C. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lapata, M. (2000). *The Acquisition and Modeling of Lexical Knowledge: A Corpus-based Investigation of Systematic Polysemy*. PhD thesis, University of Edinburgh.

Lapata, M. (2001). A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of the NAACL*.

Lapata, M., McDonald, S., and Keller, F. (1999). Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 30–36, Bergen.

Lappin, S. (Ed.). (1996). *The Handbook of contemporary semantic theory*. Oxford: Blackwell.

Larson, R. K. (1998). Events and modification in nominals. In *Proceedings from Semantics and Linguistic Theory (SALT) VIII*, Cornell University, Ithaca, N.Y.

Lee, J. and Fung, K. P. (1993). Confidence interval of the kappa coefficient by bootstrap resampling [letter]. *Psychiatry Research*, *49*(1), 97–98.

Levi, J. N. (1978). *The Syntax and semantics of complex nominals*. New York: Academic Press.

Levin, B. (1993). *English Verb Classes and Alternations a preliminary investigation*. Chicago and London: University of Chicago Press.

Levin, B. and Rappaport, M. (1986). The formation of adjectival passives. *Linguistic Inquiry*, *17*, 623–661.

Lewis, D. (1972). General semantics. In Davidson, D. and Herman, G. (Eds.), *Semantics of Natural Language*, pages 69–218. Dordrecht: Reidel.

Lombard, M., Snyder-Duch, J., and Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604.

Lui, K.-J., Cumberland, W. G., Mayer, J. A., and Eckhardt, L. (1999). Interval estimation for the intraclass correlation in dirichlet-multinomial data. *Psychometrika*, *64*(3), 355–369.

Marconi, D. (1997). *Lexical competence*. Cambridge, Mass. [etc.]: MIT Press.

Marcus, M., Santorini, B., and Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*, 313–330.

Mayol, L., Boleda, G., and Badia, T. (2005). Automatic acquisition of syntactic verb classes with basic resources. *Language Resources and Evaluation*, *39*(4), 295–312.

McCarthy, D. (2001). *PhD Thesis Lexical Acqusition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex.

McCarthy, D. and Carroll, J. (2003). Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, *29*(4), 639–654.

McDonald, R., Crammer, K., and Pereira, F. (2005). Flexible text segmentation with structured multilabel classification. In *Proceedings of HLT-EMNLP (2005)*, pages 987–994.

McNally, L. and Boleda, G. (2004). Relational adjectives as properties of kinds. In Bonami, O. and Hofherr, P. C. (Eds.), *Empirical Issues in Syntax and Semantics 5*, pages 179–196. http://www.cssp.cnrs.fr/eiss5/.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, 153–157.

Melinger, A. and Schulte im Walde, S. (2005). Evaluating the Relationships Instantiated by Semantic Associates of Verbs. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, Stresa, Italy.

Merlo, P. and Stevenson, S. (2001). Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, *27*(3), 373–408.

Mihalcea, R., Nastase, V., Chklovski, T., Tatar, D., Tufis, D., and Hristea, F. (2004). An evaluation exercise for romanian word sense disambiguation. In *Proceedings of ACL/SIGLEX Senseval-3*, Barcelona.

Miller, G. A. (1998a). Nouns in wordnet. In Fellbaum, C. (Ed.), *WordNet: an Electronic Lexical Database*, pages 23–46. London: MIT.

Miller, K. J. (1998b). Modifiers in WordNet. In Fellbaum, C. (Ed.), *WordNet: an Electronic Lexical Database*, pages 47–67. London: MIT.

Montague, R. (1974). English as a formal language. In Thomason, R. H. (Ed.), *Formal philosophy: Selected Papers of Richard Montague*, chapter 6, pages 188–221. New Haven (Conn.) (etc.): Yale University Press.

Màrquez, L., Taulé, M., Martí, M., García, M., Real, F., and Ferrés, D. (2004). Senseval-3: The catalan lexical sample task. In *Proceedings of the Senseval-3 ACL-SIGLEX Workshop*, Barcelona, Spain.

Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281.

Nirenburg, S. and Raskin, V. (2004). *Ontological Semantics*. MIT Press.

Onyshkevych, B. and Nirenburg, S. (1995). A lexicon for knowledge-based mt. *Machine Translation*, *10*(1-2), 5–57.

Parsons, T. (1970). Some problems concerning the logic of grammatical modifiers. *Synthese*, *21*(3-4), 320–324.

Partee, B. H. (1996). The development of formal semantics in linguistic theory. In Lappin, S. (Ed.), *The Handbook of Contemporary Semantic Theory*, pages 11–38. Oxford: Blackwell.

Partee, B. H. (2001). Privative adjectives: subsective plus coercion. To appear in T.E. Zimmermann, ed. Studies in Presupposition.

Passonneau, R. J. (2004). Computing reliability for coreference annotation. In *Proceedings of the Language Resources and Evaluation Conference (LREC'04)*, Lisbon.

Peters, I. and Peters, W. (2000). The treatment of adjectives in Simple: Theoretical observations. In *Proceedings of LREC 2000 (2nd International Conference on Language Resources and Evaluation)*, Atenes.

Picallo, C. (2002). L'adjectiu i el sintagma adjectival. In Solà, J. (Ed.), *Gramàtica del català contemporani*, pages 1643–1688. Barcelona: Empúries.

Poesio, M. and Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 76–83, Ann Arbor.

Popping, R. (1988). On agreement indices for nominal data. In Saris, W. E. and Gallhofer, I. N. (Eds.), *Sociometric Research, Volume 1, Data Collection and Scaling*, volume 1, pages 90–105. London: MacMillan Press.

Postal, P. (1969). Anaphoric islands. In *Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*, pages 205–239, Chicago.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge: MIT Press.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Raskin, V. and Nirenburg, S. (1996). Adjectival modification in text meaning representation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, pages 842–847, Copenhagen, Denmark.

Raskin, V. and Nirenburg, S. (1998). An applied ontological semantic microtheory of adjective meaning for natural language processing. *Machine Translation*, *13*(2-3), 135–227.

Reinberger, M. L. and Daelemans, W. (2003). Is shallow parsing useful for the unsupervised learning of semantic clusters? In Gelbukh, A. (Ed.), *Proceedings of the 4th Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2003)*, pages 304–313, Mexico City, Mexico. Springer Verlag.

Reips, U. D. (2002). Standards for internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256.

Resnik, P. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

Rigau, G., Atserias, J., and Agirre, E. (1997). Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of the eighth conference of the European chapter of the Association for Computational Linguistics*, pages 48–55, Morristown, NJ, USA. Association for Computational Linguistics.

Sanromà, R. (2003). Aspectes morfològics i sintàctics dels adjectius en català. Master's thesis, Universitat Pompeu Fabra.

Schapire, R. E. and Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. *Machine Learning*, *39*(2-3), 135–168.

Schulte im Walde, S. (2003). *Experiments on the Automatic Induction of German Semantic Verb Classes*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Schulte im Walde, S. (2006). Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, *32*(2), 159–194.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321–325.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–432.

Siegel, S. and Castelan, N. (1988). *Nonparametric statistics for the behavioral sciences* (Second ed.). New York [etc.]: McGraw-Hill.

Tastle, W. J. and Russell, J. (2003). Analysis and design: Assessing actual and desired course content. *Journal of Information Systems Education*, *14*(1), 77–90.

Thomason, R. H. (Ed.). (1974). *Formal philosophy: Selected Papers of Richard Montague*. New Haven (Conn.) (etc.): Yale University Press.

Uebersax, J. (2006). Statistical methods for rater agreement.

van Halteren, H., Daelemans, W., and Zavrel, J. (2001). Improving accuracy in word class tagging through the combination of machine learning systems. *Computational Linguistics*, *27*(2), 199–229.

van Halteren, H., Zavrel, J., and Daelemans, W. (1998). Improving data driven wordclass tagging by system combination. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, pages 491–497, Morristown, NJ, USA. Association for Computational Linguistics.

Verzani, J. (2005). *Using R for Introductory Statistics*. Boca Raton (etc.): Chapman & Hall/CRC.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference*, San Francisco. Morgan Kaufmann.

Véronis, J. (1998). A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*, pages 2–4, Herstmonceux Castle (England).

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* (Second ed.). Amsterdam [etc.]: Morgan Kaufmann.

Yallop, J., Korhonen, A., and Briscoe, T. (2005). Automatic acquisition of adjectival subcategorization from corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan.

Zernik, U. (1991). Introduction. In Zernik, U. (Ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum and Associates.